Oxford Review of Economic Policy, Volume 37, Number 3, 2021, pp. 585–617

Algorithmic fairness in credit scoring

Teresa Bono,* Karen Croxson,** and Adam Giles***

Abstract: The use of machine learning as an input into decision-making is on the rise, owing to its ability to uncover hidden patterns in large data and improve prediction accuracy. Questions have been raised, however, about the potential distributional impacts of these technologies, with one concern being that they may perpetuate or even amplify human biases from the past. Exploiting detailed credit file data for 800,000 UK borrowers, we simulate a switch from a traditional (logit) credit scoring model to ensemble machine-learning methods. We confirm that machine-learning models are more accurate overall. We also find that they do as well as the simpler traditional model on relevant fairness criteria, where these criteria pertain to overall accuracy and error rates for population subgroups defined along protected or sensitive lines (gender, race, health status, and deprivation). We do observe some differences in the way credit-scoring models perform for different subgroups, but these manifest under a traditional modelling approach and switching to machine learning neither exacerbates nor eliminates these issues. The paper discusses some of the mechanical and data factors that may contribute to statistical fairness issues in the context of credit scoring.

Keywords: ML fairness, statistical bias, penalized regression, ensemble methods

JEL classification: C55, C51, E51

I. Introduction

As society and the economy digitize and vast amounts of data are generated, the use of model-based decision-making is expanding. Firms and institutions are increasingly leveraging data and algorithms to inform and in some cases automate decisions. Machine-learning algorithms may give a predictive edge in many settings, owing to their superior ability to uncover patterns in data-rich environments, and these approaches have been used to guide decisions in settings from advertising, recruitment, employment, and

*Research carried out while an employee of the Financial Conduct Authority, UK, e-mail: teresagbono@gmail.com

**Financial Conduct Authority, UK, e-mail: karen.croxson@fca.org.uk

***Research carried out while an employee of the Financial Conduct Authority, UK, e-mail: apsgiles@gmail.com

We wish to thank Daniel Susskind and David Bholat, the editors of this special issue, and our anonymous reviewer for insightful comments and suggestions. We also thank Prudence Leung and Daniel Mittendorf for excellent assistance with data and analysis and we are grateful to Philippe Bracke and many other colleagues for very helpful discussions and input. Views expressed in this paper are those of the authors and not necessarily those of the Financial Conduct Authority. All errors remain our own.

https://doi.org/10.1093/oxrep/grab020

© The Author(s) 2021. Published by Oxford University Press.

For permissions please e-mail: journals.permissions@oup.com

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

education, through to criminal justice (see, for example, Agrawal *et al.* (2018)). In financial services, there are potentially many promising use cases. Machine-learning approaches could benefit both businesses and consumers by improving accuracy of important risk predictions in insurance and credit markets, for instance, and financial firms have begun to take advantage of the possibilities.[1]

However, the strength of machine-learning approaches in better uncovering patterns in rich consumer data has also led to concerns that they may also be better at inadvertently using sensitive information. The fear is that these techniques, some of which are less transparent than traditional linear models, may result in more accurate but less 'fair' predictions in relation to gender, race, or other protected or sensitive characteristics (e.g. Barocas and Selbst, 2016).

Socially undesirable 'bias' could arise from many sources, including the model and the data used to train this (see, for example, Mitchell *et al.* (2018)). Undesirable bias related to protected characteristics such as race or gender could arise inadvertently during model development (see Vanhoof *et al.* (2018), for instance). One factor here could be lack of diversity in the technical team itself (though simply having a diverse team is no guarantee of unbiased beliefs).[2] Equally, it could arise through deliberate actions during design and development of prediction models, or at the decision stage, when an action is taken informed by the prediction.

Importantly, though, a source of bias could simply be the data themselves where this reflects cultural bias of the past. Caliskan *et al.* (2017) provide a striking demonstration of this by showing that machine learning trained to learn word associations from ordinary human language (a standard corpus of text from the World Wide Web), replicates remarkably well a spectrum of biased associations measured by the Implicit Association Test. The machine recovers the imprints of our many human biases, including gender and racial biases (e.g. associating women with family and men with careers), without these being expressed explicitly within the training dataset.

In the context of financial services, socially undesirable bias may be embedded in previous credit decisions. This could be the case if, for example, a subgroup had unfairly been given less favourable access to credit in the past, and so ended up in default more often. This is not a hypothetical example: Dobbie *et al.* (2018) discuss the large disparities that have been observed in the availability and cost of credit across different demographic groups within many developed countries.

In this paper, we explore the potential impact of machine learning models in the context of credit scoring, a welfare-important prediction problem that affects all. Our main contribution is to compare a traditional credit scoring model to ensemble machine learning models (less interpretable but more focused on prediction) in terms of both accuracy and statistical fairness of predictions. We examine relevant statistical fairness conditions from the literature and exploit a unique dataset containing the credit files of 800,000 UK adults (a representative 2 per cent sample of the UK adult population).

A further contribution is to demonstrate the use of proxies for protected and sensitive characteristics in the context of testing for statistical fairness. The proxies we

[1] See recent surveys on machine learning in UK financial services from the Financial Conduct Authority and Bank of England: Financial Conduct Authority and Bank of England (2019); Bholat *et al.* (2020).

[2] The United Nations Gender Social Norms Index found that only 14 per cent of women and 10 per cent of men are free of bias against women (United Nations Development Programme, 2020).

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

develop are based on publicly available UK census data and thus are potentially valuable in real world settings, where those interrogating models for potential bias would typically not have access to individual-level data on many protected and sensitive characteristics. UK census data allow us to provide some insight into the fairness of credit scores in relation to subgroups that differ on protected characteristics but also on some indicators of wider vulnerability, such as poor health status and deprivation. And by using a clustering approach rather than considering sensitive characteristics individually, we are also able to take account of potentially important interactions.

We begin by reproducing a traditional linear modelling approach commonly used in practice by credit reference agencies, before constructing and implementing two variants of ensemble machine-learning methods: a form of random forest model[3] and a gradient boosted decision tree. We then focus on two questions:

(i) How does the quality of predictions of credit scoring models change as we switch from conventional statistical technology to ensemble machine learning approaches?
(ii) How does statistical fairness change? Are population subgroups that differ along protected or sensitive lines impacted differently?

On the first question, and consistent with the literature, we confirm that ensemble machine learning models deliver statistically significantly higher out-of-sample prediction accuracy than the traditional logistic model. This result holds at overall sample level as well as for individual subgroups defined along protected and sensitive lines (more on these below).

In order to answer our second question, we must first define what fairness means in this context. While 'fairness' has many philosophical and moral dimensions, here we focus on model statistical fairness in the sense of non-discrimination of decision-making models for individuals and subgroups.

A statistically fair (or non-discriminatory) credit score should reflect a person's true creditworthiness, in the sense of their likelihood to repay credit extended. Of course, credit scores designed to reflect default risk are only one element of lending decisions; other criteria, for example financial inclusion, may be considered by lenders. Note, too, that because a credit score is typically just one input into a lending decision (albeit a potentially very significant one), a statistically fair credit score could be used in an unjust or unfair lending decision.

Not all dimensions across which individuals differ should be taken into account by all models. As societies we tend to agree that there are some attributes which should not inform decision-making processes, including in the context of credit scoring. Indeed, many countries have specific laws that prohibit discrimination based on protected attributes (in the UK, the 2010 Equality Act).

Traditionally, firms have addressed these concerns by ensuring that variables for protected and sensitive characteristics are not available to models. This maps to a first very simple notion of fairness, referred to in the literature as 'fairness through unawareness'. However, this notion of fairness is too simplistic in most contexts, including credit scoring. Simply excluding information from models does not guarantee that this information

[3] We build a forest of 'extremely randomized trees'.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

is not used, as it may be somehow encoded in the data available to the model. This is particularly relevant in financial services, where decision models are very common and data on consumers is highly granular. Increased application of machine learning models could further exacerbate these issues, as these models may be better at exploiting the information contained in rich datasets. A scoring algorithm may uncover patterns correlated to a person's protected characteristics from other non-excluded data, such as credit card usage and home address (Pedreshi *et al.*, 2008). This can happen even in cases where the sensitive attribute is not highly correlated with any one feature in particular, but instead slightly correlated with many of the features. A related notion is that a simple, transparent model will lead to fairer outcomes. However, Kleinberg and Mullainathan (2018) find that simplistic models can breed unfairness and that, by increasing accuracy, more flexible algorithms can actually reduce demographic biases.

Many fairness criteria have been developed in the machine learning literature to test the statistical fairness of classifiers. We review four main notions of statistical group fairness used in classification problems (like credit scoring) in order to identify criteria relevant and practicable for the context of credit scoring.

Performance parity requires that the model be as accurate at predicting default risk for one population subgroup (e.g. females) as it is for another (e.g. males). In other words, the model must give members a fair shot at an accurate score, irrespective of their subgroup membership. Intuitively, this is a relevant property for a credit scoring model and we take this into our study as a first statistical fairness condition, while noting that a limitation with this metric is that it allows for disparities between subgroups in the pattern of errors—the same overall accuracy of credit score could be achieved in different ways for different subgroups.

Many of the statistical fairness criteria proposed in the literature can be expressed as properties of the joint distribution of the attribute, the outcome (in our case, probability of delinquency), and the score. Following Barocas *et al.* (2018), these criteria can, to a first degree approximation, be expressed by the three criteria considered here: *sufficiency*, *separation*, and *independence*

Independence requires that the model's predictions be statistically independent of the sensitive attribute (e.g. race, gender). This would require that a model generate the exact same distribution of credit scores within each population subgroup. This is not a useful fairness criterion in assessing repayment risk since we know that there are substantial income and wealth differences linked to gender and race, for example, and would expect these in turn to drive differences in actual creditworthiness. Enforcing similar credit score distributions for all subgroups would mean systematically overestimating creditworthiness for some subgroups and underestimating creditworthiness for others, and rather than being fairer, this distortion is likely to be harmful. Consider that subgroups whose creditworthiness is underestimated may be denied access to credit in cases where they could afford repayments, whereas those with credit scores that are 'too generous' may be extended credit in cases where the borrowing is unsustainable, potentially leading to harmful over-indebtedness. Hence, we exclude *independence* as a criterion for our study of fairness in credit scoring.

Separation requires that conditional on the outcome (in this case, defaulting or not), the distribution of scores is the same for each subgroup. This guarantees that for any cut-off used in a lending decision, the rate of true positives and false positives for each subgroup will be the same.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Finally, *sufficiency* requires that individuals assigned the same score by the model have the same actual risk of delinquency, irrespective of their subgroup membership. This, too, is an applicable notion for credit scoring, so *sufficiency* becomes our third criterion.

Recent work by Liu *et al.* (2019) has also shown that standard machine learning procedures implicitly aim to achieve *sufficiency* and that this will be approximately satisfied if protected characteristics can be triangulated well using other variables. However, the authors also note that a more *sufficient* risk scoring function will tend to satisfy *separation* less closely. This observed tension is consistent with a well-established theoretical result that no scoring system can satisfy exactly both *sufficiency* and *separation* simultaneously, apart from in the trivial case where the credit score perfectly predicts default or where group base rates are exactly equal (Friedler *et al.*, 2016; Chouldechova, 2017; Kleinberg *et al.*, 2017; Barocas *et al.*, 2018). However, while certain fairness measures cannot hold fully at the same time, there is evidence that some fairness measures can be correlated in practice. Friedler *et al.* (2018) make the case for avoiding a proliferation of measures in studies by focusing on a good minimal working set that includes consideration of class-sensitive error rates.

We note that our three statistical fairness criteria are observational. Observational definitions have many appealing aspects. They are convenient to work with and can be verified subject to sampling error. They do not demand knowledge of the inner workings of the model, or whether a particular feature is causally relevant for the prediction problem. Causal notions of fairness have been discussed in the literature, but practical application is often prohibitive. Truly understanding causality in the context of credit scoring is challenging, as the relationships between sensitive attributes and credit file information are everything but straightforward (see, for instance, Lee and Floridi (2020) on this point). Another limitation is that our observational criteria do not refer to the decision-maker's goals or the impact of decisions for consumer outcomes. Data limitations make it challenging to study this reliably, but understanding fairness in relation to decisions and ultimate consumer outcomes (and not just in relation to predictions) should be a priority for future work. We comment on this further below.

In order to assess our credit scoring models on the three chosen statistical fairness criteria, we must de- fine population subgroups of interest and identify these in our sample. Individual-level data on protected or sensitive characteristics are not routinely available in this prediction context (they are not typically collected by credit rating agencies or lenders, nor are they routinely available to regulators). Characteristics can, however, be proxied in some cases, subject to data and relevant consents. While access to individual-level data on actual protected and sensitive characteristics would enable a precise study of statistical fairness, if useful proxies can be created from data available to lenders and agencies in the real world, then this also has some advantages in terms of potential practical relevance of insights.

A first characteristic of interest is gender. We are able to create a fairly precise proxy for binary gender at an individual level using titles (e.g. Mr, Mrs, Ms) available in the credit file data. A caveat with this approach is that we can only consider gender binarily. Other characteristics of interest include race and potential indicators of vulnerability, such as deprivation and low health status. The latter are not themselves protected characteristics under UK law (though may correlate to some extent with protected

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

characteristics), but are independently important for welfare and of interest to policymakers. The approach we take in this paper is to offer a perspective on potential bias in relation to bundles of these attributes, developing statistical proxies using UK census data. Aggregated census data are fully publicly available, so unquestionably available to those building credit models as a potential tool to help check models for signs of demographic bias.

With census data we can identify the share of people in any small locality (i.e. an 'output area') that have particular characteristics (e.g. being white, being in poor health, being very deprived). Output areas are very granular localities that contain at least 40 households with a recommended size of 125 households. While we cannot observe protected or sensitive attributes at an individual level, the census data are granular enough to allow us to develop an indicative read on the likely presence or absence of statistical unfairness in relation to these characteristics.

In the paper, we group these localities into three clusters that differ meaningfully on the dimensions of race, deprivation score, and health status (using a K-means clustering algorithm) and then assign each individual in our dataset uniquely to one of the three clusters using their postcode. The three clusters that emerge are similar in size and reveal interesting and distinct profiles. One of the clusters is majority non-white and shows a relatively high incidence of deprivation. Another is predominantly white with a high incidence of both deprivation and poor health status. The remaining cluster is predominantly white, too, but with a low incidence of deprivation and health problems.

Overall, this means that each individual in our credit file data is assigned by us a gender (via our binary proxy) and membership of one of three geographic clusters differing on race, health status, and deprivation.[4] We then use our three credit scoring models (the traditional linear model and our ensemble machine learning approaches) to predict each individual's probability of default. In order to simulate the real-world setting for this prediction problem (including the legal obligations under which agencies operate, such as the 2010 Equality Act in the UK), we do not make any of the protected or sensitive variables available to our credit scoring models.

We find that neither traditional models nor ensemble machine learning models fully satisfy our test for statistical fairness (based on the three statistical fairness criteria above); ensemble machine learning is not obviously less fair and is statistically slightly fairer on some measures. The models predict default risk with similar overall accuracy for all of the subgroups (males, females, and, separately, our demographic clusters). We do, however, observe some small discrepancies, even with the traditional linear model. Using a linear model limits the extent to which the algorithm can uncover patterns related to the protected or sensitive characteristics we study, so this may seem unexpected. One possible factor contributing to the gaps may be differential data availability for subgroups.

Our paper is organized as follows. The next section discusses relevant literature, section III describes our dataset, and section IV introduces the credit scoring models and compares their prediction accuracy. We then shift focus to study statistical fairness: section V develops proxies for protected and sensitive char- acteristics (gender and population subgroups), and section VI evaluates the different models on statistical

[4] Data privacy is preserved throughout this classification process.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

fairness criteria, using the proxies developed in the previous section. Section VII expands our analysis to take a closer look at the relationship between the sensitive attributes and credit model training data. Finally, section VIII concludes.

II. Related literature

Our work contributes to a growing literature exploring algorithmic fairness, studying the potential for bias to arise in relation to protected characteristics such as race or gender. Over the last 50 years, many statistical criteria have been proposed to assess the fairness of both human and algorithmic decision systems. Hutchinson and Mitchell (2019) document how quantitative definitions of fairness have been developed in the context of educational testing and hiring decisions since the 1960s. More recently, research on algorithmic fairness has given rise to a multitude of related criteria. While mathematically precise, metrics are inevitably linked to value judgements about what constitutes 'fair' predictions, decisions, and outcomes. Indeed, Binns (2017) makes the point that tensions between different criteria echo longstanding debates in moral and political philosophy. Mitchell *et al.* (2018) provide an overview of approaches to assess the fairness of prediction algorithms. In this paper, we follow Barocas *et al.* (2018) in focusing on a subset of representative non-discrimination criteria. The authors note that, to a first approximation, most fairness criteria proposed fall into one of the three categories discussed above: *independence*, *separation*, *sufficiency*.

Difficulties obtaining data have generally limited opportunities for empirical study of algorithmic fairness in financial services, but there have been some insightful recent contributions. Focusing on the US con- text, some researchers have leveraged government-mandated recording of race under the Home Mortgage Disclosure Act (HMDA) and the public disclosure of HMDA data.

Fuster *et al.* (2020) analyse bias and accuracy in the context of US mortgage lending decisions, again leveraging rich HMDA data. Simulating predictions of traditional logit models and more sophisticated machine-learning approaches, they find that out-of-sample predictive accuracy is improved under machine learning for all subgroups. They show that the improved accuracy comes with increased dispersion in default probability rates, and that the expected sign of the change in estimated default propensity varies by race, with the majority (White non-Hispanic) group receiving more upgrades to their credit assessment than some minority race and ethnic groups (Black and Hispanic). They document that while the majority of the improvement comes from additional flexibility available to the machine learning approach, up to 30 per cent is attributable to triangulation of protected characteristics.

Bartlett *et al.* (2019) use HMDA to look at bias in observed lender decision-making in origination and refinance of mortgages in the US. Comparing human decisions with algorithmic lending, they find that ethnic minorities are more discriminated against in face-to-face lending: otherwise equivalent borrowers from minority groups are charged significantly higher interest rates. While algorithmic lenders do not eliminate this discrimination altogether, they do reduce the rate disparities substantially and exhibit no discrimination in rejection rates. In recent work in the UK context, Dobbie *et al.* (2018) looked at the observed lending decisions of a payday loan firm and identified significant

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

bias in lender decision-making against immigrant and older loan applicants, which they show is linked primarily to a misalignment of firm and decision-maker incentives.

Finally, another emerging area of research has begun to explore the potential impact of imposing fairness criteria on decisions. Making illustrative assumptions about lender behaviour, Hardt *et al.* (2016) estimate the reduction in profit that would result from imposing several fairness criteria when making lending decisions based on US FICO credit scores. They find that, in their stylized model, imposing a version of the independence results in the highest reduction of profit, while unawareness of protected characteristics results in the least. However, for the reasons mentioned above, both unawareness and independence are unlikely to be useful real-world fairness criteria in the context of credit scoring. Recent work by Lee and Floridi (2020) has emphasized the importance of taking into account the context and objectives of the real-world decision-maker in assessing the fairness of algorithms and proposes that any assessment of the fairness of a particular algorithm is done in relation to alternatives (rather than some absolute ideal), in a way which factors in the values of the decision-maker and considers trade-offs.

A possible consequence of imposing fairness constraints on credit-scoring algorithms is that, were lenders faced with less accurate credit scores, they may rely increasingly on 'alternative data' instead, in so far as this is permitted by regulation. Björkegren and Grissen (2019) find that predictions using data on borrowers' mobile phone history can compete or even outperform those based on conventional data available to credit referencing agencies. Liu *et al*. (2018) consider the long-term effects of imposing fairness criteria. They show that requiring fairness criteria to be satisfied has both the potential to improve and the potential to worsen long-term outcomes for disadvantaged groups.

III. Data

(i) Consumer credit rating agencies (CRA) data

We simulate credit-scoring models using large-scale credit file information sourced from credit rating agencies (CRA). Our sample contains the individual credit files for a randomly selected representative subset of the UK adult population, around 800,000 individuals, between January 2014 and December 2017. The CRA data consist of monthly snapshots of all credit holdings and current accounts. This is supplemented by information on defaults, individual voluntary arrangements (IVAs), county court judgments (CCJs), and bankruptcies, as well as information on credit product searches and applications.[5]

The CRA data also include the titles of individuals, which we use to build a binomial gender flag (see section V(i)), and postcodes, which we use to match individuals to geographic localities (using UK census data).

[5] The CRA data are described in more detail in a Technical Annex to the Financial Conduct Authority's *High-Cost Credit Review*, see: https://www.fca.org.uk/publication/feedback/fs17-02-technical-annex.pdf

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

(ii) Census data

We use public Office for National Statistics (ONS) census data in our statistical fairness tests to check how our models behave for subsets of the population that differ in relation to race, health status, and deprivation status. The data from the census are used to identify population subgroups that differ on race in combination with some indicators of vulnerability. We focus on the smallest locality available in the ONS census data, known as an output area (OA). The minimum OA size in England and Wales is 40 resident households and 100 resident people, but the recommended size is larger, at 125 households. The data cover England and Wales only and are sourced from the website of InFuse.[6] We therefore only consider individuals living in England or Wales, for whom we have census information.[7]

IV. Comparing models on accuracy

(i) The prediction problem

Consumer application credit scores estimate the creditworthiness of individual borrowers defined by the predicted probability of a future delinquency. The standard definition of a delinquency is a 'bad' outcome in credit markets, for example entering severe arrears or defaulting on a product. Below we discuss in more detail how we define delinquency.

Credit scoring models are constructed using observable characteristics derived from credit files data held by CRAs. Many countries forbid the use of protected characteristics such as gender and race in these models. In the UK, the 2010 Equality Act sets out which characteristics are not to be used and gender and race are among those attributes that must be excluded from models.[8]

We recreate a credit scoring model using a standard consumer application scoring procedure. Using the CRA data, we build:

- X_t^i, the feature set, made of 444 features describing the holdings, repayment history, search be- haviour, current account deposits, and balances of consumers in credit markets over windows of 3, 6, 9, 12, and 24 months. A list is provided in Appendix D. These are designed to reproduce the features typically used by CRAs in their models. The features do not include data on protected characteristics, such as gender or race, nor geographic data such as postcode.

X_t^i, the feature set, is defined in the t_{-23} to t_0 window for each individual i.

- a delinquency flag, Y_t^i. This is equal to 1 if one new 'bad' episode occurred over a 12-month period. A 'bad' episode is defined as one of: entering more than 90 days arrears on a product, an account opening with a debt collection firm, a new CCJ,

[6] See: http://infuse2011.ukdataservice.ac.uk/
[7] Census data for Northern Ireland and Scotland are not available in the same format.
[8] https://www.legislation.gov.uk/ukpga/2010/15/contents

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

bankruptcy, or IVA. The delinquency flag is used as the dependent variable in the credit scoring model.

$$Y_t^i = \begin{cases} 1 \textit{ if at least one 'bad' episode in the } \mathrm{t}_1 \textit{ to } \mathrm{t}_{12} \textit{ window} \\ 0 \textit{ otherwise} \end{cases} \quad (1)$$

The prediction problem can thus be expressed as finding a prediction function that maps from the credit file information to the delinquency flag. We can then use this prediction function to predict the probability of a bad episode in the period t_{13} to t_{24} for each individual i in the sample.

Most people are very unlikely to have a delinquency, and thus the distribution of the predicted probabilities returned by a credit scoring model is very skewed towards zero. To obtain a less skewed distribution, CRAs convert the predicted probabilities into interpretable scores by choosing a baseline score and odds ratio, and then normalize the scores around this point (usually with a doubling of the odds per an increase in score of a chosen size). Here, we are interested in the predicted probabilities rather than the scores.

(ii) Samples

When CRAs train their credit scoring models, they use the prediction functions they obtain to estimate the current risk of defaults of individuals. The prediction function is trained and tested on past data, for which outcomes (i.e. the delinquency flag) are observable. Because the prediction function is deployed on the most current data, for which outcomes are not yet observed, CRAs are unable to know 'how right' their predicted credit scores are, until time has passed and outcomes are observed.

Like CRAs, we train and test our models on past data. Unlike CRAs, though, we simulate deployment on past data too, so that we can evaluate how prediction functions do on more recent data than those they were trained on.

Hence, we train the models on a random 1 per cent sample of UK adults as of December 2015. This means that the dependent variable is defined between January 2016 and December 2016, and the feature set is defined between January 2014 and December 2015. We refer to these data as the training dataset. We use data for another 1 per cent sample of individuals in December 2015. We refer to these data as the test dataset.

We evaluate the prediction function against data for the individuals in the training dataset as of December 2016. We refer to this as the evaluation dataset. This means that the dependent variable in the evaluation dataset is defined between January 2017 and December 2017, and the feature set is defined between January 2015 and December 2016. This recreates the deployment environment: we obtain a prediction function by training a model on the most recent labelled data for a set of individuals, picking it with out-of-sample testing (i.e. making sure it generalizes well), and deploy it using the most recent data, which are unlabelled, for those same individuals, i.e. we produce credit scores for these individuals.

We exclude, in order: 'thin files', i.e. individuals with empty or highly incomplete credit files (32 per cent of the total), those whose postcode does not match to an OA

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

(3 per cent of remaining population), and those with missing census data (2 per cent of remaining population). All individuals have gender estimates. Thin files are a large part of the sample. Many are the result of ID duplication. For example, if an individual opens a new product shortly after changing address, they may be considered a new person and given a new ID, with no other information. This thin file would then exist until a consolidation happens, for example when the individual updates the address on one other product so that a link between the old ID and the new 'thin file' ID is created. There are other reasons why thin files exist: recent immigrants, those who are unbanked, and people without a fixed address may also have thin files. Usually CRAs use the average delinquency probability (conditional on whatever information may be available, like year of birth or postcode) to produce credit scores for those with thin files, as there are not enough data to include them in the models. Another reason is that we use a very strict definition of thin files, as we also exclude those with highly incomplete credit files. Most of these are individuals with no current account information on file, roughly 20 per cent of the overall sample. We do this because we know that almost all UK adults (around 97 per cent) have a current account,[9] meaning that the majority of these thin files are incomplete rather than truly thin.

Thin files are undoubtedly a potential source of discrimination and pose clear welfare concerns. The impact of a lack of data, in a system where access to credit is linked to data, can be great and may disproportionately affect subgroups of the population. However, for the purpose of our study, we exclude thin files as we are interested in the more subtle question of how different algorithms use the same set of complete information.

Figure 10 in Appendix A shows the data processing for training, test, and evaluation datasets. The data are very unbalanced: 5.1 per cent of people have a delinquency in the training and test datasets, 6.4 per cent in the evaluation dataset.

(iii) Accuracy of traditional model vs ensemble methods

Traditionally, credit rating agencies use linear classifiers. Therefore, for our baseline credit scoring model we fit a penalized logistic regression. We use scikit-learn LogisticRegressionCV with l2 penalization. Linear models are popular in credit scoring agencies partly because they give more interpretable results than other machine learning methods, like ensemble models or neural nets. This may come at the expense of prediction power. For the lenders using the models, even small increases in prediction power can result in large monetary gains. We simulate the switch from traditional linear models to ensemble machine learning models, to see if these models do indeed give better predictions and whether changes in prediction accuracy are similar for different population subgroups in our sample (our enquiry into statistical fairness).

To simulate the switch to more performance-focused credit scoring models, we fit two ensemble classifiers: one focused on bagging, and another on boosting. We use an extremely random forest (scikit-learn ExtraTreesClassifier) for the bagging model and gradient boosted decision trees (xgboost XGBClassifier) for the boosting model. The extremely random forest classifier is similar to the random forest, but at each node thresholds are drawn at random for the subset of features selected, and the best of these

[9] https://www.fca.org.uk/publication/consultation/cp18-42-annexes.pdf

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

random thresholds is then used for splitting. This usually reduces variance further than classic random forest models.

We tune several parameters using scikit-learn GridSearchCV, and check that the best parameters are not at the edge of the grid, in order to ensure that all models are given a chance to find the best prediction function possible. All models are tuned using a five-fold cross validation and evaluated using the area under the receiver-operating characteristic curve (AUROC).

We also fit two naïve classifiers to benchmark our models: a most frequent classifier (predicts the most common class, in our case 0, for everybody) and a stratified classifier (predicts the minority class for a random subset of observations the size of the minority class).

All five algorithms are given the same information: they all have the same features for the same individuals in the same time interval.

To evaluate the models we use two measures, firstly the area under the ROC, perhaps the most widely used evaluation metric for binary classification problems. For our second measure, given that the data are highly unbalanced, we also look at the trade-off between precision and recall; we are less interested in the ability of the model to correctly predict no delinquencies and more concerned with its ability to correctly predict delinquencies. There are many metrics to measure precision and recall in a classifier, we use the average precision (AP). AP is the weighted mean of precisions achieved at each threshold weighted by the increase in recall from the previous threshold.[10] That is:

$$\mathbf{AP} = \sum_n (R_n - R_{n-1})\, P_n. \tag{2}$$

Where R_n and P_n are, respectively, the recall and the precision at the *nth* threshold. The results are shown in Table 1 for the three models and the two dummy models.

The dummy models do as well as expected. All three models do much better than the dummy models: the AUROC are all over 0.8. While all three models perform very well, the ensemble models outperform the baseline linear model significantly: in the evaluation set, the extremely random forest closes 16 per cent of the gap in performance between the traditional linear model and a hypothetical perfect technology, while the gradient boosting model closes 20 per cent of this gap. Similarly, the ensemble models do much better than the baseline model in AP.

These results suggest that the switch to ensemble models is likely to increase the quality of predictions, giving, on average, more accurate credit scores.

V. Developing proxies for protected and sensitive characteristics

We have shown how models trained on the same feature set and data make predictions with different accuracy, suggesting that different algorithms are able to extract different information from the same data.

[10] We prefer AP to the area under the precision recall curve as AP does not require interpolation. The area under the precision-recall curve is estimated using the trapezoidal rule, which uses linear interpolation and can be too optimistic. See Boyd *et al.* (2013).

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Table 1: Performance of the three models and the two dummy models

Dataset	Model	AUROC	Precision-recall (AP)
Training	logit	0.878	0.528
	etc	0.939	0.636
	xgb	0.928	0.673
	frequency	0.500	0.064
	stratified	0.500	0.064
Test	logit	0.875	0.514
	etc	0.895	0.539
	xgb	0.900	0.549
	frequency	0.500	0.063
	stratified	0.500	0.063
Evaluation	logit	0.888	0.585
	etc	0.906	0.606
	xgb	0.910	0.616
	frequency	0.500	0.073
	stratified	0.500	0.073

A reasonable question to ask at this point is whether particular models perform differently for different subsets of the population. In other words, are the results of the models statistically 'fair', or do some exhibit bias in their predictions towards or against certain subgroups? In order to study this, we must define subgroups of the population that differ meaningfully on characteristics of concern. We focus in this paper on gender and race, both protected characteristics under the UK Equality Act, and two indicators of vulnerability: health status and deprivation.

(i) Proxying gender

Individual-level data on protected or sensitive characteristics are not routinely available to regulators, nor are they routinely collected by firms. However, the CRA data has individual-level data on titles (e.g. Mr, Mrs, Ms, Miss) as reported at the time of taking out a credit product. We use this information to estimate a binary gender proxy. This proxy works well for the vast majority of cases. Where it is not conclusive on gender, for example for military or medical titles, we randomly assign gender using the gender split for that title at UK population level. One notable limitation with this approach is that we can only consider gender binarily.

(ii) Clustering localities on race and vulnerability

To uncover subgroups of individuals in the sample that differ meaningfully on characteristics of interest, we cluster the 175,434 output areas in England and Wales on race, health status, and deprivation using census data, and then assign each individual in our sample to one of these clusters based on the person's postcode.

The motivation behind clustering the local areas by characteristics is two-fold. First, these characteristics tend to be strongly correlated at the level of a local area, and investigating each in turn would likely have revealed similar conclusions with respect to each characteristic. Second, the literature on statistical fairness with respect to group

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

membership is much more developed than the literature on fairness with respect to continuous characteristics.[11] In keeping with this literature, the set of characteristics is based on identifying a privileged group in that dimension, and then calculating the share of the local area population that are not in that privileged group. This approach has the undesirable property that it glosses over potentially relevant differences between those who are not members of the privileged group—for example, the causes of poor health, or individuals' actual race or ethnicity. But this is required to make the problem statistically manageable, given the data available to us.

For clustering variables we use: proportion of individuals with poor health status, proportion of deprived households, and proportion of non-white individuals. These are defined straightforwardly from the raw census variables as follows:

- The proportion of individuals with poor health status in an OA is defined as the sum of those with bad health and those with very bad health over the output area total population.
- The proportion of deprived households in an OA is defined as the number of households deprived in three or four dimensions over the total number of households in a output area.[12]
- The proportion of white people in an OA is defined as the sum of people self-identifying as white in an output area over the total number of people in an output area.

Race is a protected characteristic under the UK 2010 Equality Act. Health status and deprivation, while potentially correlated to some extent with protected characteristics like 'disability', are not protected under the Act. Both do, however, relate to vulnerability, and fair treatment of vulnerable consumers is an important focus for policymakers. As discussed above, an attractive aspect of census data is that they are fully publicly available, so something anyone building models might in principle be able to use as part of a strategy to check their model for potential biases.

Clusters

We cluster OAs in England and Wales on the dimensions of race, deprivation score, and health status, using a K-means clustering algorithm. We choose to use three clusters, having considered both the silhouette coefficients and the within sum of squares at various clusters. The clusters identified are meaningfully different across the three dimensions:

- **cluster 0** 'non-white and deprived'—majority BAME subgroup, relatively deprived, with medium incidence of health problems;
- **cluster 1** 'white and relatively well-off'—predominantly white with low household deprivation and low incidence of health problems;
- **cluster 2** 'white, severely deprived, poorer health'—predominantly white with high incidence of health problems and household deprivation.

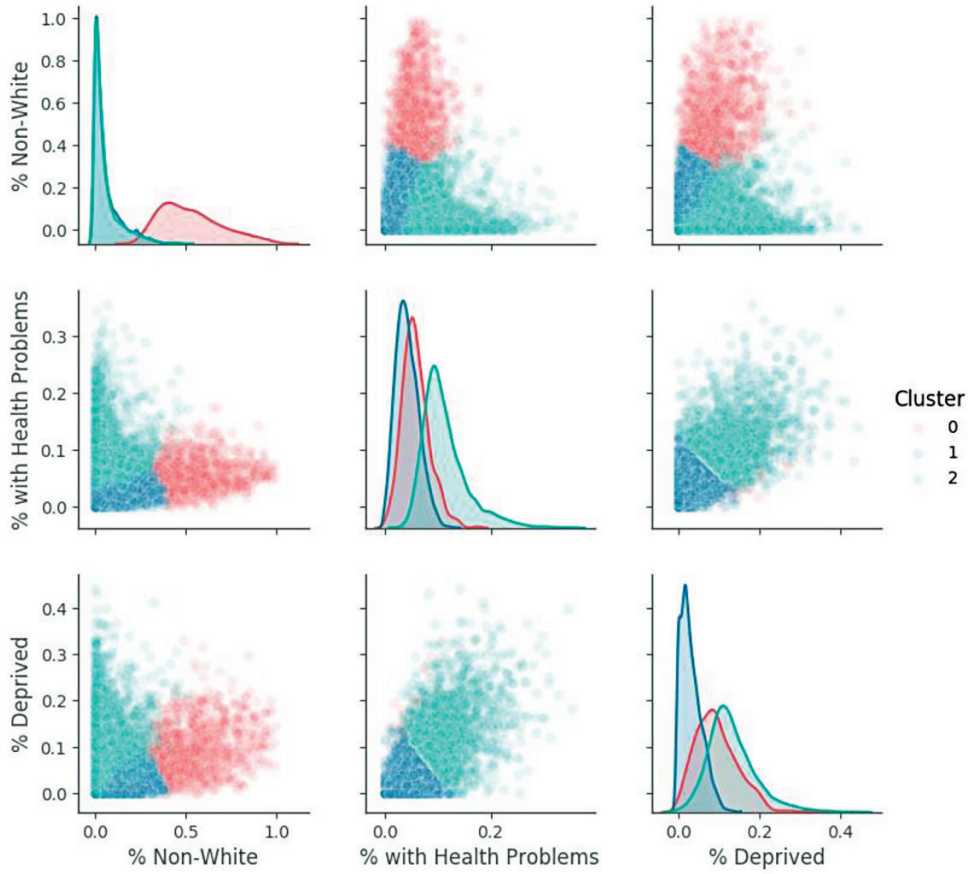
[11] For example, age. In principle many of the tests applied in this paper can be adapted to a continuous characteristic, but it complicates the analysis and exposition.

[12] https://www.statistics.digitalresources.jisc.ac.uk/dataset/classification-household-deprivation-great-britain-2011

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Figure 1 illustrates the clusters visually.

Figure 1: Summary of race/vulnerability clusters



Notes: The diagonal shows the distribution for each cluster variable by cluster. Scatter plots show each pair of cluster variables for each cluster. Each individual geographic locality (output area) is represented by a dot. Cluster labels are: **cluster 0** 'non-white and deprived'; **cluster 1** 'white and relatively well-off'; **cluster 2** 'white, severely deprived, poorer health'.

We checked the robustness of our clustering by testing the stability of the algorithm and by comparing the results to the clusters obtained by running the k-means algorithm on the principal components obtained applying principal component analysis of the raw census variables. See Appendix C for further details and robustness checks.

VI. Comparing models on statistical fairness

The next step is to apply statistical fairness criteria to the predictions of the different credit scoring models, focusing on fairness in relation to gender (using our proxy) and subgroups of the population that differ on race and vulnerability (our clusters).

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

As discussed in our introduction, many statistical fairness criteria have been put forward in the literature, but not all definitions are meaningful in all contexts. Reviewing the main notions of statistical fairness (see Barocas *et al.* (2018)) and screening for relevance and applicability in the credit scoring context, we focus on *performance parity*, *separation*, and *sufficiency*.

Performance parity requires equally good prediction power for each population subgroup. In the credit scoring context, it requires that the credit scoring model be as accurate at predicting delinquencies for one population subgroup as it is for another (e.g. female or male). Intuitively, large disparities in accuracy may signal fairness issues, in the same way that in the gender gap literature the difference in average wage between men and women is often used as a starting point to signal the presence of underlying fairness issues. While this is a meaningful criterion to consider, it is in no way definitive: it does not tell us anything on the causes of such differences.

Separation requires that conditional on the outcome (in this case, defaulting or not), the distribution of scores is the same for each subgroup. This guarantees that for any cut-off used in a lending decision, the rate of true positives and false positives for each subgroup will be the same.

Sufficiency means that individuals assigned the same score by the model should have the same true risk of default, irrespective of their subgroup membership.

(i) Performance parity

Examining our first notion of statistical fairness, we see overall model performance (measured as AUROC) is somewhat lower for males and less privileged groups under the traditional scoring approach (see Tables 2 and 3 and also Figures 2 and 3). Ensemble models neither close the gap between subgroups nor exacerbate it. Interestingly, with the ensemble models, AUROC for the 'white and deprived' cluster increases to the point that the gap vs 'white and relatively well-off' is eliminated. However, those belonging to the 'white, deprived, and poor health' cluster continue to receive lower average prediction performance compared to other subgroups.

It is important to stress that this statistical fairness concept is different from broader concepts of fairness. Increased quality of prediction for a subgroup may mean that ensemble models are better at 'separating' one subgroup from another. Both subgroups now have more 'truthful' scores, i.e. scores that are closer to their true creditworthiness, but this may mean that one group now has lower credit scores than before. As a society, we may deem it more fair to pool the risk of specific subgroups, even if this means less accurate scores.

What could explain the differences we see even in the traditional model? There are a few potential factors—not mutually exclusive. One is the lower representation of some subgroups in the available data: the model learns only from the data it sees and there are naturally fewer data points for some subgroups (e.g. ethnic minorities) in the population. Fewer data mean less opportunity to learn.

Another possibility is that the prediction problem itself is intrinsically easier for some subgroups than for others, because their average default risk is particularly low or particularly high, say. Suppose the model observes two subgroups with the same frequency (i.e. data are not the issue), but that people in one camp are very low risk, while the

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Table 2: Performance of the three models and the two dummy models by gender

Dataset	Metric	Gender	Logit	Etc	Xgb	Frequency	Stratified
Training	aucs	Female	0.889	0.947	0.936	0.500	0.499
		Male	0.867	0.931	0.920	0.500	0.502
	prs	Female	0.554	0.662	0.697	0.065	0.064
		Male	0.500	0.608	0.648	0.064	0.064
Test	aucs	Female	0.886	0.905	0.910	0.500	0.500
		Male	0.864	0.885	0.890	0.500	0.501
	prs	Female	0.539	0.565	0.574	0.064	0.064
		Male	0.486	0.511	0.522	0.063	0.063
Evaluation	aucs	Female	0.897	0.914	0.918	0.500	0.500
		Male	0.879	0.898	0.903	0.500	0.501
	prs	Female	0.604	0.624	0.634	0.072	0.072
		Male	0.566	0.588	0.598	0.074	0.074

Table 3: Performance of the three models and the two dummy models by cluster

Dataset	Metric	Cluster	Logit	Etc	Xgb	Frequency	Stratified
Training	aucs	0 Non-white and deprived	0.849	0.919	0.902	0.500	0.501
		1 White and relatively well-off	0.884	0.949	0.936	0.500	0.500
		2 White, severely deprived, poorer health	0.863	0.919	0.912	0.500	0.500
	prs	0 Non-white and deprived	0.491	0.597	0.626	0.074	0.075
		1 White and relatively well-off	0.513	0.625	0.667	0.048	0.048
		2 White, severely deprived, poorer health	0.580	0.676	0.711	0.116	0.116
Test	aucs	0 Non-white and deprived	0.845	0.866	0.869	0.500	0.501
		1 White and relatively well-off	0.881	0.901	0.907	0.500	0.499
		2 White, severely deprived, poorer health	0.863	0.881	0.886	0.500	0.502
	prs	0 Non-white and deprived	0.468	0.490	0.500	0.073	0.073
		1 White and relatively well-off	0.496	0.524	0.533	0.048	0.048
		2 White, severely deprived, poorer health	0.572	0.591	0.602	0.113	0.114
Evaluation	aucs	0 Non-white and deprived	0.856	0.879	0.884	0.500	0.501
		1 White and relatively well-off	0.894	0.912	0.916	0.500	0.501
		2 White, severely deprived, poorer health	0.874	0.891	0.895	0.500	0.498
	prs	0 Non-white and deprived	0.555	0.577	0.586	0.090	0.090
		1 White and relatively well-off	0.570	0.593	0.602	0.054	0.054
		2 White, severely deprived, poorer health	0.632	0.646	0.657	0.129	0.129

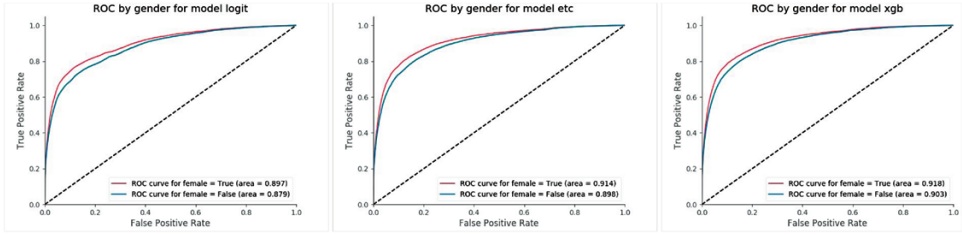
riskiness of those in the other has a strong random component, so that based on the information we have they are as likely to repay as they are to default. The nature of the problem means that the first group is easier to predict.

(ii) Separation

Turning to *separation*, we find that focusing on individuals who go on to default, men and those belonging to the white and relatively well-off cluster typically have lower predicted probability of default than other groups, i.e. the traditional model seems more 'generous' towards risky individuals drawn from these subgroups, assigning them a lower delinquency risk. When we switch to the ensemble models this 'male effect' remains: male defaulters continue to be treated more generously, other things equal.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

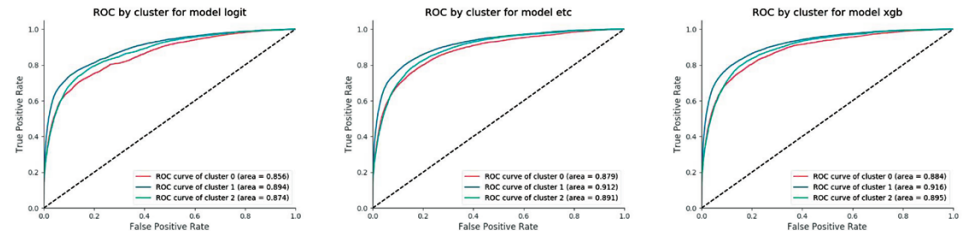
Figure 2: Model performance (AUROC) for evaluation dataset by gender



(a) Logit (b) Extra trees (c) Gradient Boosting

Note: The dotted line represents an AUROC of 0.5. The red line and blue line depict model AUROC for females and males, respectively.

Figure 3: Model performance (AUROC) for evaluation dataset by race/vulnerability cluster



(a) Logit (b) Extra tree (c) Gradient boosting

Note: The dotted line represents an AUROC of 0.5. Cluster labels are: **cluster 0** 'non-white and deprived'; **cluster 1** 'white and relatively well-off'; **cluster 2** 'white, severely deprived, poorer health'.

The plots in Figure 4 and in Figure 5 show the density function of the predicted probability of delinquency for those subsequently observed to have a delinquency. Simply put, it shows the distribution of $\hat{y}$ for those with $y = 1$.

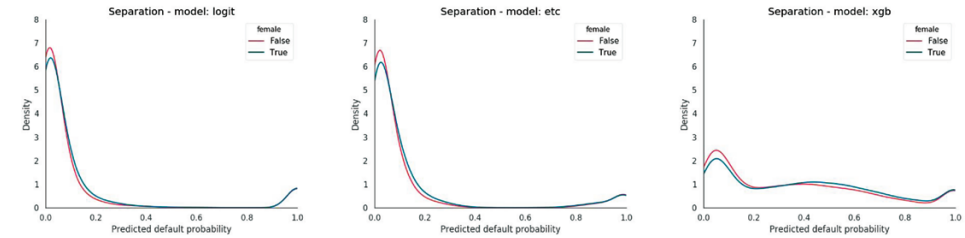
We can also look at the cumulative plots. As for the density plots discussed, the cumulative distribution function (cdf) plots in Figures 6 and 7 show very clearly how the distribution of predicted delinquencies, even for the subset of delinquents, is concentrated around zero with another small mass around 1. Estimating the delinquency risk of those on the verge of delinquency and those with very good credit files is easy, but a good credit scoring model is one that can also distinguish close cases around the middle of the distribution. The extra gradient boosting model is better than the other two models at this, as we can see from the cumulative and density plots for delinquents.

(iii) Sufficiency

Finally, *sufficiency* seems satisfied for all three models: comparing by gender and sensitive cluster, people at the same predicted default probability have the same observed probability of defaulting, on average. Put differently, when we look at a particular model it seems equally well calibrated for the different population subgroups. However, we note that the *sufficiency* criterion is satisfied 'mechanically' in cases where

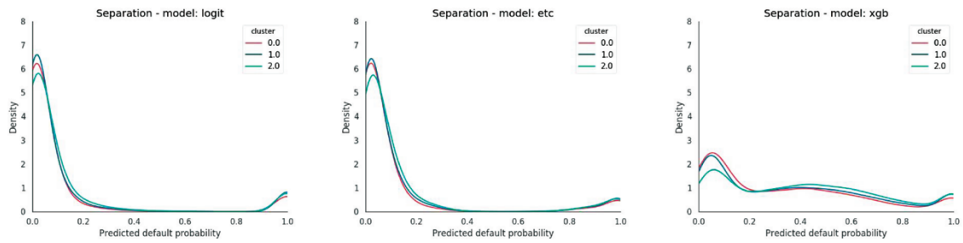
Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Figure 4: Separation by gender for delinquents.



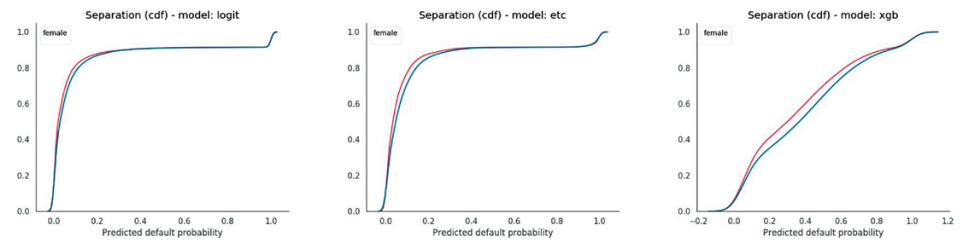
(a) Logit (b) Extra tree (c) Gradient boosting

Figure 5: Separation by race/vulnerability cluster for the delinquents



(a) Logit (b) Extra tree (c) Gradient boosting

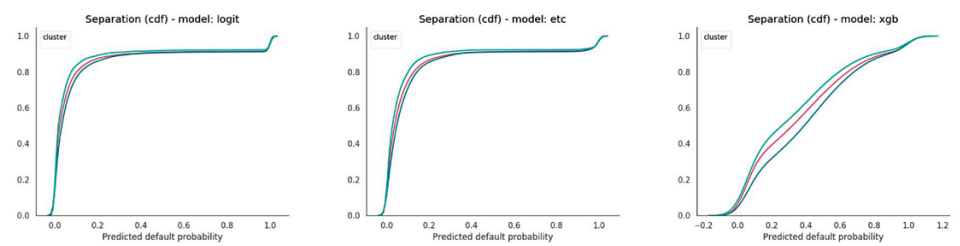
Figure 6: Separation by gender for the delinquents—cdf



(a) Logit (b) Extra tree (c) Gradient boosting

Note: The plot shows the cumulative distribution of estimated delinquency probability for those who become delinquent (i.e. for which *y* is 1) by gender.

Figure 7: Separation by race/vulnerability cluster for the delinquents—cdf

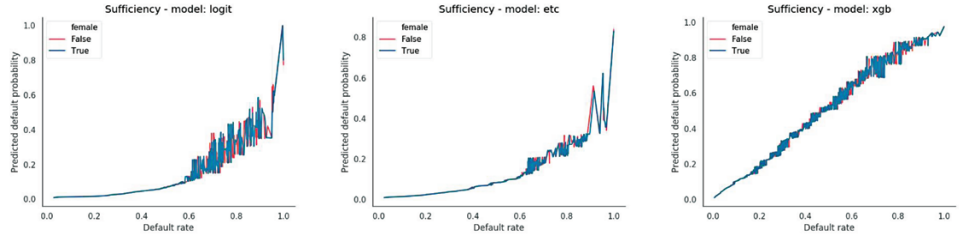


(a) Logit (b) Extra tree (c) Gradient boosting

Note: The plot shows the cumulative distribution of estimated delinquency probability for those who become delinquent (i.e. for which *y* is 1) by race/vulnerability cluster.

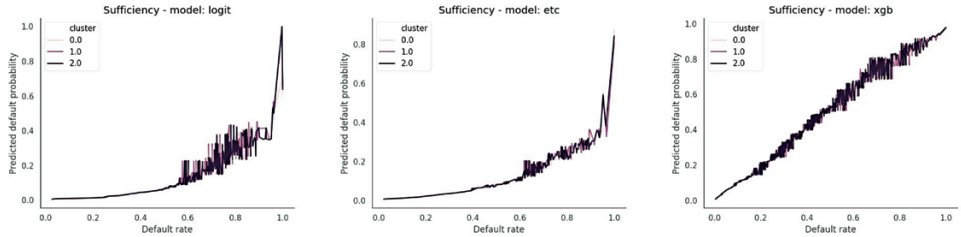
Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Figure 8: Sufficiency by gender



Note: The plot shows the rolling mean realized delinquency for each predicted probability of delinquency, by gender.

Figure 9: Sufficiency by race/vulnerability cluster



Note: The plot shows the rolling mean realized delinquency for each predicted probability of delinquency, by cluster.

the characteristics are fully encoded in the data (i.e. even when the protected or sensitive attributes are not included explicitly in the model, but enough information about them is contained in other, non-sensitive variables used in the modelling).

Figures 8 and 9 show the predicted probability of delinquency at each delinquency rate by gender and cluster. We can see that both for genders and clusters the relationships are very similar. The plots are noisier in the middle of the distribution, where the are very few people, and less noisy around 1 and 0, where most of the density of the distribution is concentrated.

VII. Relationship between sensitive characteristics and credit scoring data

The observational fairness criteria suggest a nuanced story. There are some differences across genders and subgroups, both in terms of overall prediction accuracy and through the lenses of the *separation* criterion, though we are unable to uncover any causality. We also find that *sufficiency* is satisfied. To explore this further, we expand our analysis in two ways: first, we train new credit scoring models, this time including the sensitive attributes as features. Second, we train new models using the gender proxy

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

and the subgroup membership as outcomes, and the standard credit scoring model variables as features.

(i) What changes if sensitive characteristics are included directly in the model?

It is interesting to ask how the performance of our credit models would change if we included explicitly in the training data information on the gender and ethnicity/vulnerability subgroup of individuals (something not done in the real world). Are the sensitive personal characteristics we study in this paper informative for delinquency risk beyond the information already contained in real world data sets used in credit scoring? We find that predictions improve, but only marginally. Table 4 shows the performance of the three models re-trained to include the gender proxy and the subgroup membership flag as features. The AUROC and the AP increase for all three models, but only marginally.

One possible explanation for this result is that the relevant information carried by these new sensitive features is already 'encoded' in the original feature set, so making them explicit to the models does not give them much additional information. This is in line with what we observed in the *sufficiency* plots. Among the three models, the sensitive features appear to give proportionally more information to the logit model, which is the model we would expect to be less effective in decoding latent information from the feature set. Another possible explanation is that the two extra variables do not carry relevant new information for the task at hand (predicting delinquencies), and thus the models' performance does not change substantially when the sensitive features are added.

(ii) Can sensitive characteristics be predicted from standard credit model training data?

If the gender and subgroup membership information is somehow encoded in the original features we use to train our credit scoring models, we would expect to be able to train models to predict the gender and subgroup information using these same features. However, we find that this is only partially true. In order to test this, we train another six models using the original features (i.e. excluding the gender and subgroup membership): three binary classifiers using the gender proxy as target variable, and three

Table 4: Performance of the three credit scoring models with information on gender and ethnicity/vulnerability clusters included in the training feature set

Dataset	Model	AUROC both	AP both	AUROC	AP
Training	logit	0.879	0.529	0.878	0.528
	etc	0.955	0.673	0.939	0.636
	xgb	0.926	0.660	0.928	0.673
Test	logit	0.876	0.514	0.875	0.514
	etc	0.896	0.540	0.895	0.539
	xgb	0.901	0.549	0.900	0.549
Evaluation	logit	0.889	0.586	0.888	0.585
	etc	0.907	0.608	0.906	0.606
	xgb	0.911	0.617	0.910	0.616

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurariaibrary user on 17 June 2024

Table 5: Performance of the models trained to estimate gender (proxied) from standard credit model training data

Dataset	Model	AUROC
Training	logit	0.566
	etc	0.810
	xgb	0.654
Test	logit	0.564
	etc	0.604
	xgb	0.613
Evaluation	logit	0.562
	etc	0.605
	xgb	0.612

Table 6: Performance of models trained to estimate race/vulnerability cluster membership from standard credit model training data

Dataset	Model	AUROC
Training	logit	0.620
	etc	0.768
	xgb	0.686
Test	logit	0.618
	etc	0.645
	xgb	0.649
Evaluation	logit	0.620
	etc	0.647
	xgb	0.651

multi-class classifiers using the subgroup membership as outcome. Results suggests that the models can predict gender and race/vulnerability cluster membership better than at random, but not greatly so (see Tables 5 and 6).

Interestingly, gender is harder to predict than cluster membership. This may be explained by how important these attributes are to the original prediction problem: gender may be less informative for credit scoring models than information about race and vulnerability. We also find that ensemble models are better at predicting both gender and race/vulnerability subgroup membership, which is in line with our first finding that these models are better at predicting default for a given set of information.

AUROC for the subgroup membership multi-class classifiers is the average of the AUROC for each of the three subgroups, weighted by their frequency.

Overall, these results pose more questions than they answer. We observe differences in the performance of credit scoring models by gender and by subgroups that differ on race and indicators of vulnerability, even though the training data exclude these protected and sensitive characteristics. This is true whether we simulate a traditional (linear) credit scoring model or models based on more sophisticated ensemble machine learning methods. *Sufficiency* is satisfied by all models, which would suggest the gender and subgroup information is already encoded in the original feature set. Indeed, the models do not become significantly better at predicting delinquency when the gender and race/vulnerability cluster information is included in the feature set. Both modelling approaches fail to satisfy the *separation* condition fully, which is to be expected in light

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

of the finding that each satisfies *sufficiency* and the impossibility result regarding simultaneous complete satisfaction of these two conditions.

Lastly, when we take the original feature set and train our models on this to predict gender and race/vulnerability clusters, we find that the resulting classifiers are not much better than chance. This suggests that overall the sensitive information in our gender and cluster data is not hugely important to the prediction of delinquency, and most of its predictive power is likely already encoded in the original feature set derived from standard credit files.

VIII. Conclusion

Our work provides a first large-scale empirical study of the accuracy and statistical fairness of different credit scoring technologies in the UK context; we compare a traditional logit model with more flexible machine learning approaches known for their predictive power in data-rich environments. We confirm that machine learning methods give an overall predictive edge out-of-sample, and that the statistical fairness implications appear potentially quite nuanced. None of the credit score modelling approaches we simulate entirely satisfies statistical fairness in relation to population subgroups defined along protected and sensitive lines. Even with the traditional model, there are some differences in the way the model performs for subgroups vs an ideal that would be fully statistically fair. Our hypothetical switch to machine learning appears to improve some aspects of statistical fairness, but results overall are subtle; machine learning technology neither eliminates nor appears to exacerbate the statistical fairness issues we detect.

We run two further exercises, in order to learn more about the relationship between sensitive characteristics and credit file data. In the first of these, we find that none of the models predicts delinquency significantly better when sensitive information (gender and demographic cluster membership) is included directly in the feature set. This is in line with our finding that the statistical fairness condition of *sufficiency* is met, as this happens naturally when the sensitive attributes are encoded in the features. In the second exercise, we check whether it is possible to use the original features alone to predict the sensitive attributes. The resulting classifiers are better than random at this, we find, but not greatly so. Overall, our evidence suggests that gender and demographic cluster membership are somehow correlated with information relevant for the default prediction problem (i.e. the credit scoring model), but that most of the relevant information they carry is already 'encoded' in the original credit file data.

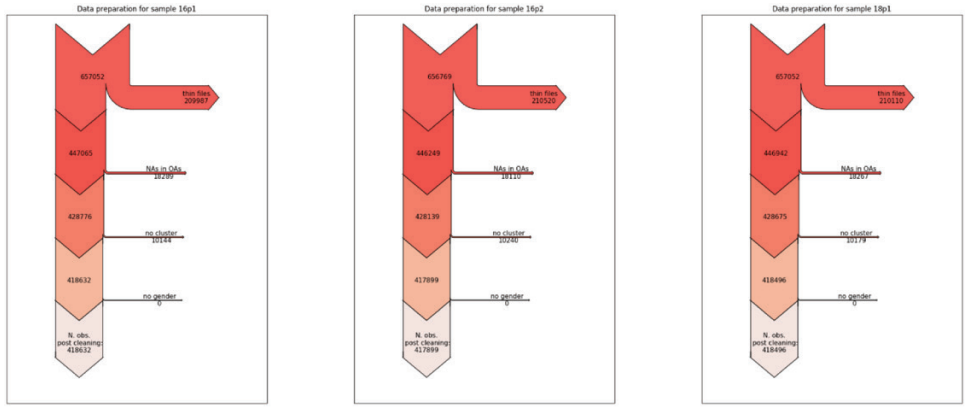
Our work is just a starting point on many fronts. It is first of all observational, and as such does not tell us anything definitively about the mechanisms that may be causing the statistical fairness gaps we observe. Further, our work does not take into account how changes in the credit scoring models may impact real- world lending and pricing decisions, and hence ultimate customer welfare. Barocas *et al.* (2018) argue that purely statistical criteria are not sufficient to prove the presence or absence of discrimination—this will depend how lending decisions are actually made. This echoes the economic perspective on algorithmic fairness provided by Cowgill and Tucker (2020), who suggest that effective regulation should focus on outcomes, not 'engineering practices or technologies'. More research is needed on the implications of alternative scoring technologies for ultimate consumer outcomes, considering how different approaches (and

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

any fairness constraints imposed) may impact lenders' actions and feed through into equilibrium pricing and access decisions. Understanding statistical fairness of credit model predictions (the focus of this paper) is then one part of a potentially much wider consideration of overall fairness in relation to the decisions of lenders and ultimate outcomes for consumers.

Appendix A: Data cleaning pipeline

Figure 10: Cleaning pipeline



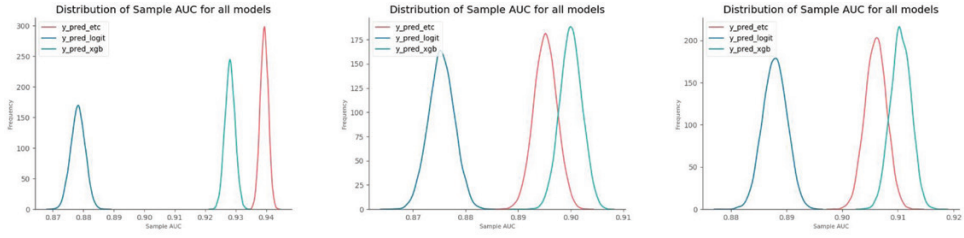
(a) Training dataset (b) Test dataset (c) Evaluation dataset

Appendix B: Bootstrapping

Here we present results from the bootstrapping of AUROC, carried out to ensure that the differences we find in the paper are unlikely to be due to chance (see Figures 11, 12, and 13). We sample 25 per cent of the individuals with replacement 10,000 times to form the evaluation set.

Figure 14 shows results from bootstrapping of the *separation* condition. We sample 25 per cent of the individuals with replacement 10,000 times to form the evaluation set.

Figure 11: Bootstrapping AUROC of credit scoring models

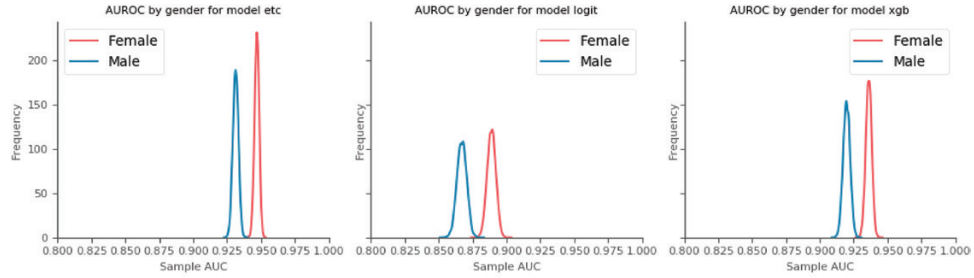


(a) Train (b) Test (c) Evaluation

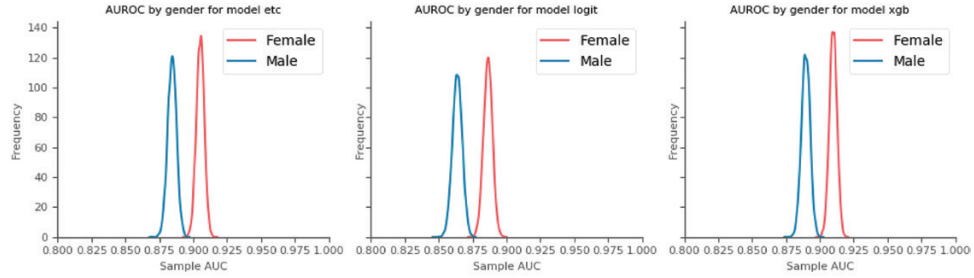
Note: Plots show the distributions of AUROCs obtained in the bootstrapping samples.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

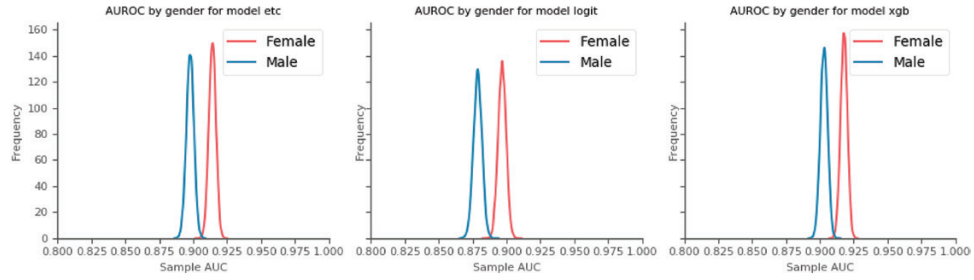
Figure 12: Bootstrapping AUROC of credit scoring models by gender



(a) Train



(b) Test

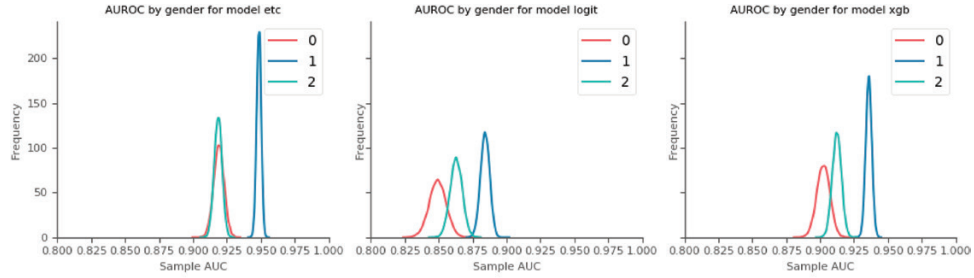


(c) Evaluation

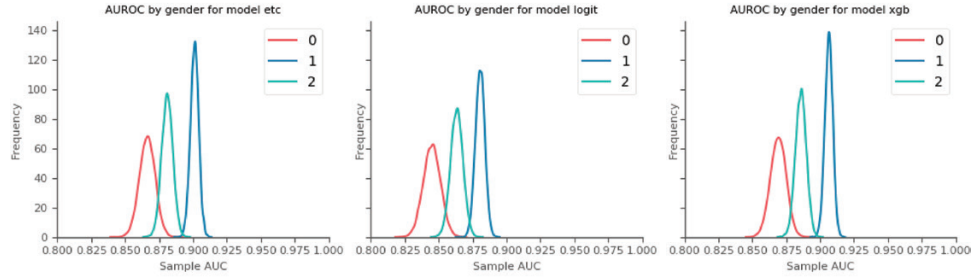
Note: Plots show the distributions of AUROCs obtained in the bootstrapping samples.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

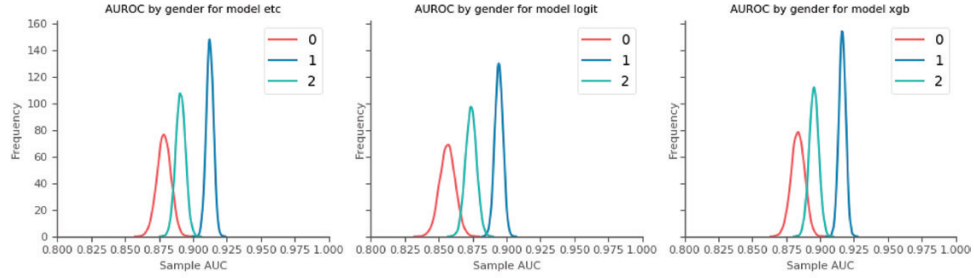
Figure 13: Bootstrapping AUROC of credit scoring models by race/vulnerability cluster



(a) Train



(b) Test

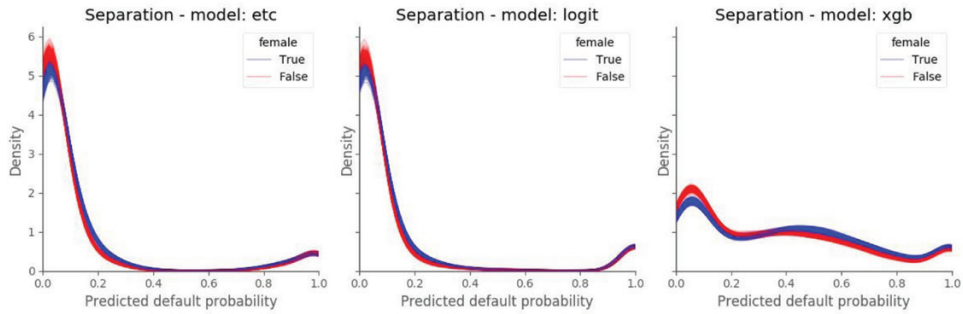


(c) Evaluation

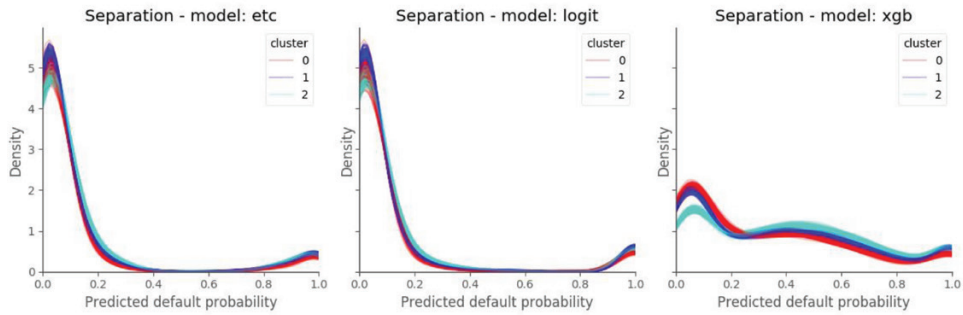
Note: Plots show the distributions of AUROCs obtained in the bootstrapping samples.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Figure 14: Bootstrapping of the *separation* condition for delinquents in the evaluation set



(a) Gender



(b) Race/vulnerability cluster

Note: Plots show the distributions of AUROCs obtained in the bootstrapping samples.

Appendix C: Clustering on race and vulnerability

Table 7 and Figures 15 and 16 present further details and robustness checks on our clustering of localities (output areas) by race, health status, and deprivation scores, including our choice of k (number of clusters).

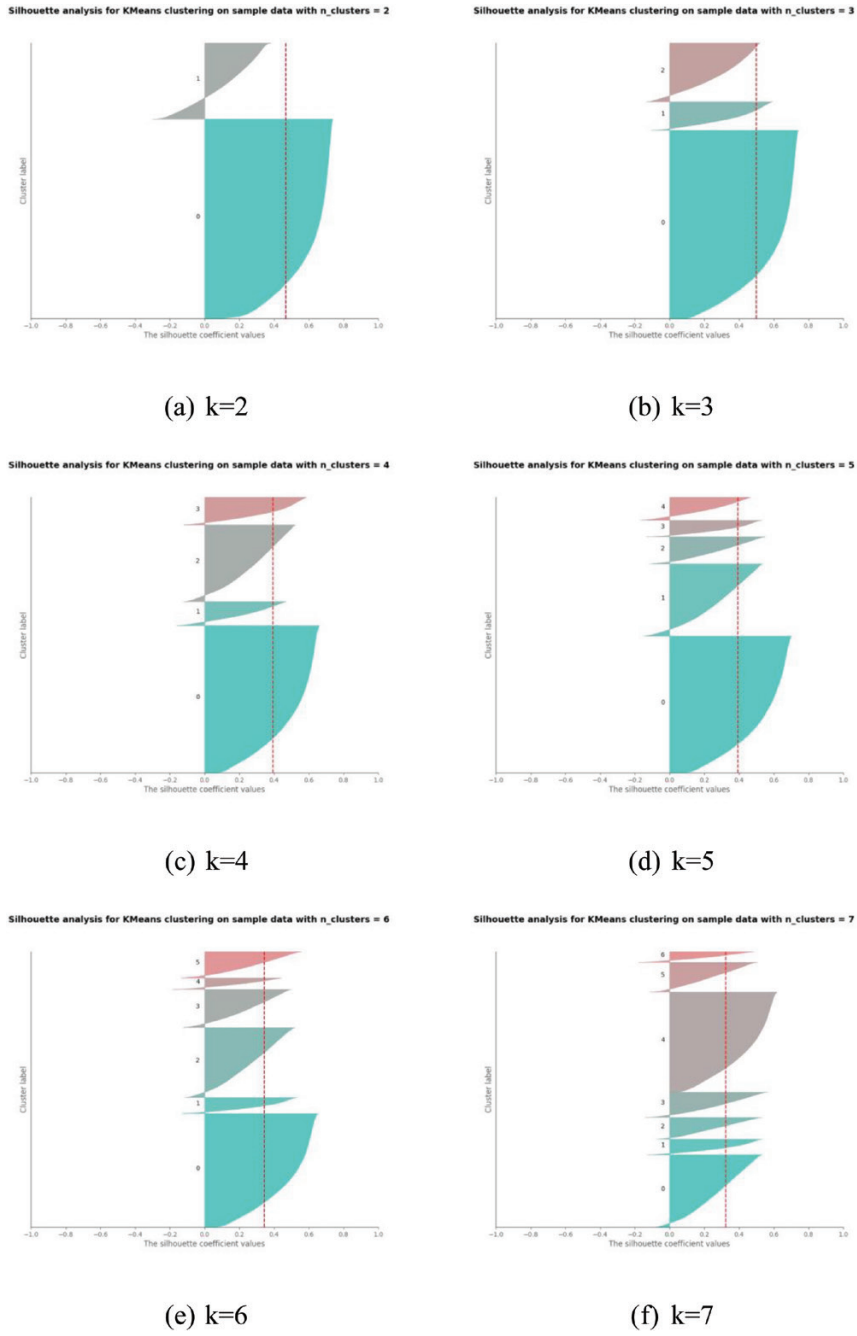
Figure 17 shows the average composition of the three race/vulnerability clusters we identify in the paper, along the dimensions used to define them: race, health status, deprivation.

Table 7: Silhouette coefficients for k between 2 and 7

	2	3	4	5	6	7
Silhouette coeff.	0.469	0.499	0.394	0.394	0.345	0.324

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

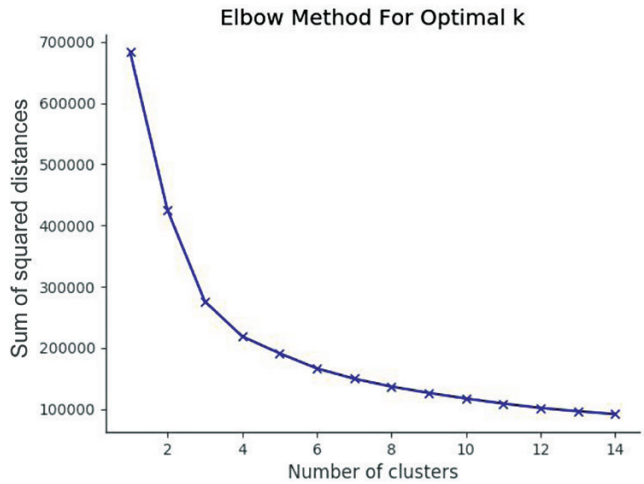
Figure 15: Silhouette plots



Notes: Silhouette analysis is a method of interpretation and validation of consistency within clusters of data. It provides a graphical representation of how well each individual has been classified. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters. This coefficient has a range of (–1, 1), where a high value indicates that the object is well matched to its own cluster and poorly matched to neighbouring clusters. See: https://scikit-learn. org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html. The vertical line shown is the average silhouette score.

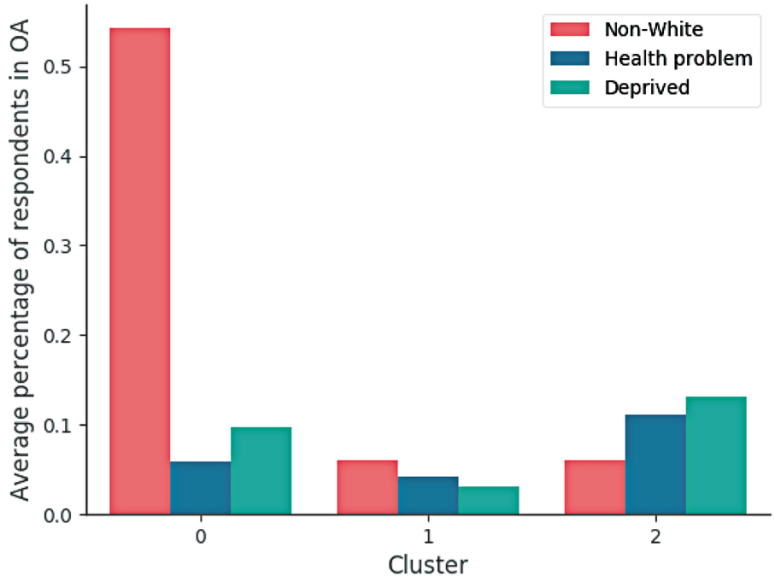
Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Figure 16: Elbow method results for k 2 to 14, where k is the potential number of clusters



Note: The plot shows the sum of squared distances to the nearest cluster centre for each k.

Figure 17: Average prevalence of each clustering variable within the output areas (OAs) assigned to each race/vulnerability cluster (for example, the average non-white population in the OAs in cluster 0 is 55 per cent)



Appendix D: Features

Table 8 catalogues all variables used in our models in the paper. For each variable, we constructed features rolling backwards by 3, 6, 9, 12, and 24 months. For models in section VII(i) we also included the gender proxy and the race/vulnerability cluster variable as features.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

Table 8: List of all variables using in the credit scoring models

all cards max payment	all prod total credit limit	currents worst status	run mean utilization max
all cards max payment max	all prod total credit limit max	currents worst status max	run mean utilization mean
all cards mean utilization	all prod total debt	currents youngest account age	run tot bal
all cards mean utilization max	all prod total debt max	fixed tot bal	run tot bal max
all cards mean utilization mean	all prod total debt mean	fixed tot bal max	run tot bal mean
all cards n cards	all prod worst status	fixed tot lim	run tot lim
all cards n cards max	all prod worst status max	fixed tot lim max	run tot lim max
all cards n cashad	all prod youngest account age	max ccj	run tot lim mean
all cards n cashad sum	all prod youngest account age max	max ccj sat	run total utilization
all cards tot bal	bal	max ccj sat sum	run total utilization max
all cards tot bal max	bal delta	max ccj sum	run total utilization mean
all cards tot lim	bal max	missing cato	short term n accounts
all cards tot lim max	bal mean	missing cato sum	short term tot bal
all cards tot payments	bal med	mortgage debt	short term tot bal max
all cards tot payments max	bal min	mortgage debt max	short term tot open bal
all cards total utilization	bal sd	mortgage in 90d arrears	short term tot open bal max
all cards val cashad	bal sum	mortgage in 90d arrears max	short term worst running status
all cards val cashad sum	cc bal in arrears	mortgage in arrears	short term worst running status max
all loans n accounts	cc best status	mortgage in arrears max	storecard n accounts
all loans n accounts max	cc best status max	mortgage in default	storecard n accounts max
all loans n accounts mean	cc max bal	mortgage in default max	storecard worst running status
all loans tot balance	cc max bal max	mortgage max bal	storecard worst running status max
all loans tot balance max	cc max limit	mortgage n accounts	thin file
all loans tot balance mean	cc max limit max	mortgage paid off	total ccj amt
all prod bal in 90d arrears	cc mean utilization	mortgage paid off max	total ccj amt sat
all prod bal in 90d arrears max	cc min bal	mortgage tot payments	total ccj amt sat sum
all prod bal in arrears	cc min bal max	mortgage worst status	total ccj amt sum
all prod bal in arrears max	cc n cards	mortgage worst status max	turnover
all prod bal in default	cc n cards max	mortgage youngest account age	turnover max
all prod bal in default max	cc n in arrears	n ccj	turnover mean
all prod best status	cc n limit increase	n ccj sat	turnover med
all prod best status max	cc n limit increase sum	n ccj sat sum	turnover min
all prod mean account age	cc n over limit	n ccj sum	turnover sum
all prod mean account age max	cc payment	n firms	
all prod n accounts	cc payment max	n firms max	

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurariaLibrary user on 17 June 2024

Table 8: Continued

all cards max payment	all prod total credit limit	currents worst status	run mean utilization max
all prod n accounts max	cc payment mean	n income searches	
all prod n closed accounts	cc tot bal	n income searches sum	
all prod n closed accounts sum	cc tot bal max	new bai	
all prod n firms	cc tot bal mean	new bai sum	
all prod n firms max	cc tot lim	repays n missed repays	
all prod n in 90d arrears	cc tot lim max	repays n missed repays sum	
all prod n in 90d arrears max	cc total utilization	repays n repays	
all prod n in arrears	cc total utilization max	repays n repays max	
all prod n in arrears max	cc total utilization mean	repays n repays mean	
all prod n in default	cc worst status	repays total missed repays	
all prod n in default max	cc worst status max	repays total missed repays sum	
all prod n new accounts	currents n accounts	repays total repays	
all prod n new accounts sum	currents n accounts max	repays total repays max	
all prod oldest account age	currents oldest account age	repays total repays mean	

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by auraraillibrary user on 17 June 2024

References

Agrawal, A., Gans, J., and Goldfarb, A. (2018), *Prediction Machines: The Simple Economics of Artificial Intelligence*, Cambridge, MA, Harvard Business Review Press.

Barocas, S., and Selbst, A. D. (2016), 'Big Data's Disparate Impact', *104 California Law Review 671*, doi: 10.2139/ssrn.2477899.

— Hardt, M., and Narayanan, A. (2018), *Fairness and Machine Learning*, fairmlbook.org, http://www.fairmlbook.org.

Bartlett, R. P., Morse, A., Stanton, R., and Wallace, N. (2019), 'Consumer-lending Discrimination in the Fintech Era', NBER Working Paper No. w25943.

Bholat, D., Gharbawi, M., and Thew, O. (2020), 'The Impact of Covid on Machine Learning and Data Science in UK Banking', *Bank of England Quarterly Bulletin Q4*.

Binns, R. (2017), 'Fairness in Machine Learning: Lessons from Political Philosophy', *CoRR*, abs/1712.03586, http://arxiv.org/abs/1712.03586.

Björkegren, D., and Grissen, D. (2019), 'Behavior Revealed in Mobile Phone Usage Predicts Credit Repayment', *The World Bank Economic Review*, doi: 10.1093/wber/lhz006.

Boyd, K., Eng, K. H., and Page, C. D. (2013), 'Area Under the Precision-recall Curve: Point Estimates and Confidence Intervals', in H. Blockeel, K. Kersting, S. Nijssen, and F. Železný (eds), *Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg, Springer, 451–66.

Caliskan, A., Bryson, J. J., and Narayanan, A. (2017), 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases', *Science*, **356**(6334), 183–6.

Chouldechova, A. (2017), 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments', *Big Data*, **5**.

Cowgill, B., and Tucker, C. (2020), 'Algorithmic Fairness and Economics', *Columbia Business School Research Paper*, doi: 10.2139/ssrn.3361280.

Dobbie, W., Liberman, A., Paravisini, D., and Pathania, V. (2018), 'Measuring Bias in Consumer Lending', NBER Working Paper No. w24953.

Financial Conduct Authority and Bank of England (2019), 'Machine Learning in UK Financial Services', *FCA Research Notes*, https://www.fca.org.uk/publication/research/research-note-on-machine-learning-in-uk-financial-services.pdf

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016), 'On the (Im)possibility of Fairness', *arXiv*:1609.07236.

— — — Choudhary, S., Hamilton, E. P., and Roth, D. (2018), 'A Comparative Study of Fairness-enhancing Interventions in Machine Learning', *arXiv*:1802.04422.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2020), 'Predictably Unequal? The Effects of Machine Learning on Credit Markets', Working Paper, doi: 10.2139/ssrn.3072038.

Hardt, M., Price, E., and Srebro, N. (2016), 'Equality of Opportunity in Supervised Learning', *Advances in Neural Information Processing Systems 29 (NIPS)*.

Hutchinson, B., and Mitchell, M. (2019), '50 Years of Test (Un)fairness: Lessons for Machine Learning', FAT19: Proceedings of the Conference on Fairness, Accountability, and Transparency, doi: 10.1145/ssrn.3287560.

Kleinberg, J. M., and Mullainathan, S. (2018), 'Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability', *Computing Research Repository*, abs/1809.04578.

— — Raghavan, M. (2017), 'Inherent Trade-offs in the Fair Determination of Risk Scores', Proc. 8th ITCS.

Lee, M. S. A., and Floridi, L. (2020), 'Algorithmic Fairness in Mortgage Lending: From Absolute Conditions to Relational Trade-offs', *Minds and Machines*, doi: 10.2139/ssrn.3559407.

Liu, L. T., Simchowitz, M., and Hardt, M. (2019), 'The Implicit Fairness Criterion of Unconstrained Learning', *arXiv*:1808.10013.

Mitchell, S., Potash, E., and Barocas, S. (2018), 'Prediction-based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions', *ArXiV*: 1811.07867v2.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008), 'Discrimination-aware Data Mining', in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 560–8, New York, NY.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024

United Nations Development Programme (2020), 'Tackling Social Norms', *Human Development Perspectives*, http://hdr.undp.org/sites/default/files/hd_perspectives_gsni.pdf

Vanhoof, M., Reis, F., Ploetz, T., and Smoreda, Z. (2018), 'Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics', *Journal of Official Statistics*, **34**(4), doi: 10.2478/jos-2018-0046.

Downloaded from https://academic.oup.com/oxrep/article/37/3/585/6374682 by aurarialibrary user on 17 June 2024