The Fairness of Credit Scoring Models*

Christophe Hurlin† Christophe Pérignon‡ Sébastien Saurin§

October 29, 2021

Abstract

In credit markets, screening algorithms discriminate between good-type and bad-type borrowers. This is their raison d'être. However, by doing so, they also often discriminate between individuals sharing a protected attribute (e.g. gender, age, racial origin) and the rest of the population. In this paper, we show how to test (1) whether there exists a statistically significant difference in terms of rejection rates or interest rates, called lack of fairness, between protected and unprotected groups and (2) whether this difference is only due to credit worthiness. When condition (2) is not met, the screening algorithm does not comply with the fair-lending principle and can be qualified as illegal. Our framework provides guidance on how algorithmic fairness can be monitored by lenders, controlled by their regulators, and improved for the benefit of protected groups.

Keywords: Discrimination; Credit markets; Machine Learning; Artificial intelligence

JEL classification: G21, G29, C10, C38, C55.

*We are grateful to seminar participants at the Credit Research Centre (University of Edinburgh), CREST and HEC Paris, as well as participants to the 2021 Hi! Paris workshop on AI bias and Data Privacy, 2021 Risk Forum, and to the Online Corporate Finance Seminar for their comments. We thank the ACPR Chair in Regulation and Systemic Risk, the Fintech Chair at Dauphine-PSL University, and the French National Research Agency (Ecodec ANR-11-LABX-0047, F-STAR ANR-17-CE26-0007-01, CaliBank ANR-19-CE26-0002-02) for supporting our research. All the results displayed in this paper has been reproduced by cascad.

†University of Orléans, LEO, Rue de Blois, 45067 Orléans, France. Email: christophe.hurlin@univ-orleans.fr

‡HEC Paris, 1 Rue de la Libération, 78350 Jouy-en-Josas, France. Email: perignon@hec.fr

§University of Orléans, LEO, Rue de Blois, 45067 Orléans, France. Email: sebastien.saurin@univ-orleans.fr

1 Introduction

For their proponents, the growing use of Artificial Intelligence (AI) in credit markets allows processing massive quantity of data using powerful algorithms, hence improving classification between good-type and bad-type borrowers.[1] The resulting lower forecasting errors correspond to fewer non-performing loans and less money left on the table for algorithmic lenders. Furthermore, credit scoring algorithms permit to include small borrowers traditionally overlooked by standard screening techniques used with standard data (Berg et al. (2020a)) and lead to higher responsiveness of the credit supply to demand shocks and lower interest rates (Fuster et al. (2019)).

However, the development of AI has also stirred a passionate debate about the associated potential discrimination biases (O'Neil (2017), Bartlett et al. (2021a)). Indeed, when automatically assessing the creditworthiness of loan applicants, credit scoring models can place groups of individuals sharing a protected attribute, such as gender, age, citizenship or racial origin, at a systematic disadvantage. The latter can be in terms of either rejection rate or interest rate. For instance, the Apple Pay app was publicly criticized for setting credit limits for female users at a much lower level than for otherwise comparable male users (Vigdor (2020)). Using detailed administrative data on US mortgages, Fuster et al. (2021) find that the use of machine-learning algorithm increases interest rate disparity between White/Asian borrowers and Black/Hispanic borrowers. When it arises, such difference is not only detrimental for the groups being unfavourably treated, it is also a potential source concerns for algorithmic lenders as it leads to severe reputation risk and legal risk. Indeed, under U.S. fair-lending law, lenders can only discriminate for creditworthiness reasons (Morse and Pence (2020)).

However in practice, how can we know whether a credit-scoring algorithm discriminates against a protected group only for creditworthiness? This question is particularly challenging to address in the context of opaque, black-box algorithms analyzing thousands of features for each borrower, covering credit and job history, transaction-level banking data, credit card data, social-media posts or other forms of digital footprint, etc. In this paper, we design and implement a methodology allowing algorithmic lenders, as well as their regulators, to address this question. We proceed in three steps. First, we quantify the difference of treatment across groups and formally test whether we can reject the null hypothesis of equality of treatment, which we call fairness. Treatment can

[1] In this paper, we use AI and Machine Learning (ML thereafter) interchangeably to describe algorithms able to learn by identifying relationships within data and to produce predictive models in an autonomous manner.

be considered either in terms of access to credit (acceptance rate) or price (interest rate). Second, we propose a novel and simple interpretability technique, called *Fairness Partial Dependence Plot* (FPDP), to identify the variables that cause the lack of fairness. Third, once the candidate variables have been identified, we check whether all these variables are legitimate in the lending context, namely that they meet a legitimate business need that cannot reasonably be achieved otherwise. Importantly, when some variables appear not being legitimate in the lending context, the algorithm is not complying with the fair-lending principle and the resulting decisions could be challenged in court.

In our analysis, we focus on several definitions of fairness which appear particularly relevant in the lending context. The most commonly used definition, *statistical parity*, corresponds to the equality of probability of being classified as good type in all groups, i.e., those displaying the protected attribute vs. those not displaying the protected attribute. We also make use of *conditional statistical parity*, which states that the probability of being classified as good type conditional on both displaying the protected attribute (e.g. being a woman) and belonging to a given homogenous risk class (e.g. range of income, range of job tenure, married, and no past credit event) is the same in all groups. An advantage of this conditional version is to control for potential composition effects across groups. Furthermore, in order to control for classification errors, we also consider definitions that condition on the type of borrowers.

In practice, algorithmic lending can lead to significant differences in access to credit across groups for various reasons. First, credit scoring algorithms can be trained on a dataset gathering past decisions made by biased loan officers. In this case, the algorithms inherit and perpetuate human biases in their decision-making. Alternatively, the scoring algorithm can learn from a dataset of actual defaults that occurred in the past. Two cases arise: *disparate treatment* occurs when the algorithm explicitly uses the protected attribute and *disparate impact* occurs when the lender use facially neutral variables that are able, collectively, to synthetically reconstruct the protected attribute (Fuster et al. (2021),Bartlett et al. (2021a), Prince and Schwarcz (2020), Bellamy et al. (2019)). In practice, disparate impact is more likely to occur with highly non-linear, non-interpretable, and opaque scoring models, such as advanced ML algorithms.

Our methodology fits nicely within the US legal framework to ensure fairness in lending, and in particular the Equal Credit Opportunity Act (ECOA) and Fair Housing Act (FHA), see Evans (2017), Evans and Miller (2019), and Bartlett et al. (2021b). Under this framework, the plaintiffs making a claim of unintentional discrimination must demonstrate that a lending practice impacted disparately on members of a protected group. This corresponds to our first step, namely testing the null hypothesis of equality of treatment. If disparate impact has been shown, the framework then demands that the burden shifts to the defendant to show that the practice is consistent with business necessity. This is exactly the purpose of our second and third steps: identifying the variables that cause the lack of fairness and checking their legitimacy in the lending context. Thus, our method is a way to operationalize statistically the legal concept of fair lending for any scoring model.

We illustrate our fairness assessing framework by testing for gender discrimination in a database of retail borrowers. First, we show that our tests are able to detect direct discrimination when gender is explicitly used as a feature to assess creditworthiness. Second, when gender is not included in the feature space, most considered scoring models turn out to be fair regardless of the considered metrics. Interestingly, the null hypothesis of fairness is still rejected for some highly non-linear, flexible ML models. Furthermore, we show that the choice of the parameters controlling the learning process of the algorithms strongly impacts fairness. In models displaying a lack of fairness, our interpretability technique identifies the set of candidate features. Reassuringly, the selected features make sense as they play a key role in the algorithm's decision rules. When scrutinizing candidates variables, we discover two types of variables: (1) those who appear to be legitimate: they correlate with both gender and default, and they have theoretical reasons to help forecasting default, and (2) those who appear non-legitimate: exhibiting lower or no correlation with gender and default, as well as no particular theoretical reasons to do so. We interpret the presence of less legitimate variables as an indication that the algorithm may not comply with the fair-lending principle. Finally, we show how to increase the fairness of the model, while controling for performance.

Making sure that AI algorithms treat individuals, and especially bank customers, in a fair way is nowadays a top priority for governments and regulators, as demonstrated by recent reports and white papers devoted to the governance of AI in finance. For instance, the proposal for a regulation of AI released by the European Commission in April 2021 states that "*AI systems used to evaluate the credit score or creditworthiness of natural persons should be classified as high-risk AI systems [...] AI systems used for this purpose may lead to discrimination of persons or groups and perpetuate*

historical patterns of discrimination, for example based on racial or ethnic origins, disabilities, age, sexual orientation, or create new forms of discriminatory impacts".[2] The potential discriminatory biases of AI in financial services are also receiving a lot of attention by international media, think-tanks, and consulting firms.[3] However, while the issue is now universally recognized, academics, lenders, regulators, customer protection groups, lawers and judges are still lacking tools to look at it in a fair and systematic way. The resulting high legal and regulatory uncertainties surrounding the use of ML algorithms acts as an impediment for financial service providers to innovate and invest in screening technologies (Evans (2017), Bartlett et al. (2021b)). We see our paper as an attempt to fill this gap.

We make several contributions to the literature on discrimination in lending. First, we propose a standalone methodology to formally check whether a given credit scoring model complies with the fair-lending principle. Beyong its academic value, it can be directly used in practice. For instance, it could help algorithmic lenders to monitor the fairness of their models, or be used by their regulators to set guidelines or control models. Our method could also prove handy for lawyers and judges to conduct legal expertise. While we share the same objective of identifying illegitimate lending practice, our approach differs in some important ways with the legal and economic framework of Bartlett et al. (2021b). First, they focus on the input used by the algorithm whereas we develop a backtesting approach focusing on algorithms' outcomes. Second, our framework relies on the concept of algorithmic fairness, which has become a central concept for scholars working on algorithmic discrimination in machine learning (Barocas et al. (2020)), law (Gillis and Spiess (2018)), and business and economics (Kleinberg et al. (2018), Cowgill and Tucker (2020)).

Our second contribution is to propose the first inference tests for fairness definitions. This formal approach offers several advantages: First, it allows us to assess fairness while taking into account estimation risk. Fairness assessment is fundamentaly based on the comparison of conditional probabilities (e.g., probabilities of being correctly classified by the scoring model) accross different groups, which have to be estimated. Thus, controling for estimation risk is crucial to avoid misconclusions about the fairness of the scoring model. Second, it can advantageously replace ad-hoc rules used in practice, and in court, to decide whether decisions in different groups of individuals are "similar

[2]Other recent examples include the recent reports published by US (Lael (2021)) or international regulators (European Commission (EC (2020)), European Banking Authority (EBA (2020)), or French authority of banking and insurance supervision (ACPR (2020)).

[3]For media coverage, see for instance Anselm (2020): "Is an Algorithm Less Racist Than a Loan Officer?" - The New York Times. For discussion by think-tanks, see the coverage by the Brookings institute, Klein (2020). For coverage by consulting firms, see Deloitte (2020) and PWC (2021).

enough". An example of such rules is the four-fiths rule, which states that a group is discriminated against if and only if the rate of being favorably classified in this group is less than 80% of the rate in the rest of population.

Our third contribution is to show how to improve the fairness of machine-learning algorithms. To do so, we develop a novel interpretability method allowing us to identify the variables with the strongest impact on a fairness metric. We show how to treat some of these variables to reduce non-legitimate differences between groups of applicants, yet maintaining a high level of performance.

The rest of our paper is structured as follows. We review the literature and legal framework on discrimination in lending in Section 2 and we present several fairness definitions which are relevant in the context of lending in Section 3. Section 4 details our common framework for testing for discrimination biases in credit models. We illustrate our methodology in Section 5 using a dataset of retail loans. Finally, we conclude our study in Section 6.

2 Discrimination in Lending

Discrimination in lending across demographic groups can take two main forms: higher rejection rate and higher interest rates. Furthermore, it can happen at any stage of the life of a credit: when applying for a new loan, when refinancing it, or when asking for a credit limit extension. Several standard economic theories can explain this phenomenon. Under taste-based discrimination (Becker (1957)), some managers get utility from engaging in discrimination against individuals sharing a protected attribute, and even so when these individuals are more productive. In this setting, discriminating firms are then not maximizing profit as they pass on economically-attractive opportunities. Differently, under statistical discrimination (Arrow (1971), Phelps (1972)), firms lack information about the true creditworthiness of borrowers. One way to deal with this uncertainty is to rely on the average historical creditworthiness of each group of borrowers, where groups rely on protected attributes or other features. Alternatively, firms can rely on variables that are correlated with both creditworthiness and a protected attribute.

There is compeling empirical evidence about discrimination in lending. Bartlett et al. (2021a) show that hispanic and African-American borrowers pay 7.9 and 3.6 basis points more in interest for home-purchase and refinance mortgages, respectively, because of discrimination (see also Bhutta and Hizmo (2021)). These higher price tags represent 11.5% of lenders' average profit per loan. As their identification strategy neutralises the effect of creditworthiness, the authors can attribute

this price difference to discrimination. Similarly, Bayer et al. (2018) find that after conditioning on credit characteristics, African American and Hispanic borrowers were 103% and 78% more likely, respectively, than other borrowers to be in a high-cost mortgage.[4] D'Acunto et al. (2021) show, in the context of an Indian peer-to-peer platform, that individual lenders from a given religious group are reluctant to lend to the members of another religion and, at the same time, implement lax-screening to borrowers from their religion. In her study of the pictures posted on the prosper.com lending platform, Ravina (2019) finds that good-looking borrowers are more likely to get a loan and more likely to default. Dobbie et al. (2018) study the long-term profits made by a high-cost lender in the UK. They find that immigrant (respectively older) applicants yield long-term profits that are nearly four (two) times larger than native-born (younger) applicants. Conversely, they report no bias against female applicants. Their findings suggest that these differences are mainly due to loan officers' incentives, in line with Berg et al. (2020b).

The legal framework for discrimination in lending corresponds to the fair lending laws (Evans and Miller (2019), Bartlett et al. (2021b), Bartlett et al. (2021a)). In the US, the Federal Reserve, along with other consumer protection agencies, enforces two federal laws that ensure fairness in lending: ECOA and FHA. The former applies to both consumer and commercial credit and prohibits credit discrimination on the basis of racial origin, gender, color, age, national origin, marital status, or receipt of income from any public assistance program. FHA applies to credit related to housing and overlaps extensively with ECOA in terms of protected attributes, with the exception of handicap, which is specific to FHA. Under these lending laws, two situations are unlawful: (1) disparate treatment occurs when the lender treats a borrower differently because of a protected attribute and (2) disparate impact occurs when the lender use facialy neutral variables that, both, adversely affect the members of a protected group and do not meet a legitimate business need that cannot be reasonably achieved otherwise. It is important to note that both situations remain unlawful even if there is no conscious intent to discriminate. There are similar fair lending laws in most jurisdications outside the US, although they are included in broader anti-discrimination laws (e.g. UK and Canada). The European Union recognizes non-discrimination as a fundamental right as shown in Article 21 of the EU Charter of Fundamental Rights: *"Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited"*. However, each member state can develop its own legislation

[4]Other price differences are reported for women in Italy (Alesina et al. (2013)).

and enforcement policy. The resulting differences across countries make the EU anti-discrimination regulation a less clear and coherent normative anchor than ECOA and FHA.

The rise of algorithms and big data in lending can have an important impact on the likelihood and forms of discrimination. First, using an algorithm making objective decisions and applying same standards to all customers can reduce or even remove taste-based discrimination (Philippon (2019)). Supportive evidence is provided by D'Acunto et al. (2021) who show that robo-advising appears to fully debias lenders by equating the share of borrowers of each religion in lenders' portfolios to the share of borrowers of each religion on the Indian P2P platform they investigate. Similarly, Bartlett et al. (2021a) find that the discrepancy between the rates charged to White and Black borrowers is lower for algorithmic lenders than conventional lenders and that the former exhibit no disparities in mortgage rejection rates. Furthermore, shifting to algorithmic lending can also aleviate the pernicious effect of misaligned incentives as shown theoretically and empirically by Dobbie et al. (2018).

Second, ML algorithms, and especially when implemented with large datasets, are likely to better capture the structural relationship between observable characteristics and default (Jansen et al. (2020)). Here the effect on discrimination can go both ways. On the one hand, as shown by Berg et al. (2020a), combining advanced modeling techniques with non-standard data permits to include small borrowers traditionally overlooked by standard screening techniques. Conceptually, ML can also give access to credit to people who are credit invisible because of being unbanked or because they lack credit history. On the other hand, the model and empirical evidence in Fuster et al. (2021) indicate that ML increases rate disparity across groups of borrowers and benefits more White and Asian borrowers than Black and Hispanic borrowers. Compared to standard parametric scoring models (e.g. logistic regression), ML models introduce an additional flexibility which improves out-of-sample classification accuracy. However, the gains associated to this improvement are not homogeneously distributed accross borrowers, as minorities may be affected by a triangulation effect. The latter occurs when non-linear associations between the features proxy for the protected variable, hence "de-anonymizing" the group identities only using non-protected attributes.[5]

Third, algorithms can reflect and perpetuate human biases if the training of the algorithm is made on past human decisions or if the training set lacks diversity (e.g. fewer women paying back their loans in due time). This problem is particularly acute when the training set is not

[5]Triangulation is also labelled proxy discrimination, disparate impact, or indirect discrimination (Prince and Schwarcz (2020)).

representative of the entire population (see Dastin (2018) for a discussion of Amazon's infamous automated CV screening device). Fourth, algorithm can accomodate the use of behavioral data or digital footprints, which can exacerbate lending discrimination. As exlained by Evans (2017), in Federal Trade Commision (FTC) vs. CompuCredit, the FTC alleged that the lender's scoring model penalized consumers for using their cards for certain types of transactions, such as paying for marriage counseling or therapy. Furthermore, a Department of Justice enforcement action involved a lender whose models excluded borrowers with a Spanish-language preference.

How can we show that a given lending practice, algorithmic or not, is illegal because of discrimination? As explained by Bartlett et al. (2021b), there is a long tradition in the US of applying the burden-shifting framework to cases of statistical discrimination. Under this framework, plaintiffs must prove that a particular lending practice affects negatively and significantly the members of a protected group. If this is the case, the defendant must show that the practice is consistent with business necessity. In case, the defendant is not able to provide compelling evidence that this is indeed the case, the lending practice is said to be illegitimate, and then illegal.

Consistent with this framework, Bartlett et al. (2021b) formulate a statistical test to apply to the design and review of the *inputs* used in any algorithmic decision-making processes. Their test, called input accountability test, seeks to exclude variables that are correlated with both default and the protected attribute. An alternative approach is to rely on the *output* or outcome of the algorithm. According to Cowgill and Tucker (2020), outcome-based approaches are preferable as they exhibit more flexibility, fewer loopholes, greater efficiency, and stronger incentives for innovation. One common outcome-based approach is to rely on a fairness definition. A critic often addressed to this approach is that there are multiple definitions of fairness, and that several of them are incompatible with one another (Berk et al. (2021)). As a result, using such approach requires a pre-specified level of discrimination that is permissible in the outcomes (Gillis and Spiess (2018)). In practice, the level of tolerance is set arbitrarily. For instance, the Uniform Guidelines on Employee Selection Procedures, adopted in 1978 by the Equal Employment Opportunity Commission (EEOC (1978)) introduces the 80% rule as follows "*a selection rate for any racial origin, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact.*" Differently, below we propose an output-based method in which the

level of tolerance is given by statistical theory.

3 Measuring fairness

3.1 Framework and notations

We consider a bank using an algorithm to screen borrowing applications. The binary output label Y takes values in $\mathcal{Y} = \{1, 0\}$ where the value 1 corresponds by convention to the favorable output.[6] In the context of credit scoring, examples of favorable outcome include loan-application approval, refinancing approval and overdraft authorization, etc. For convenience, we will thereafter use the *good type* ($Y = 1$) vs. *bad type* ($Y = 0$) terminology. The vector X denotes the non-protected features, which include borrower features (e.g., income, assets, debt-to-income ratio, age, occupation, banking and payment data) and contract terms (e.g. loan size, loan-to-value ratio). We denote by D the *sensitive* or *protected* attribute (e.g. racial origin, gender, age, religion). In the following, we model the protected attribute as a binary variable, $D = \{0, 1\}$, where the value 1 refers to an applicant belonging to the protected group and 0 an applicant belonging to the unprotected group.[7]

The goal of the bank is to build a scoring model which maps the observable non-protected features X into a conditional probability for an applicant of being good type, $p(X) = \Pr(Y = 1|X)$. This probability is then transformed into a predicted outcome $\widehat{Y}$ taking a value 1 when $p(X)$ is above a given threshold $\delta \in]0, 1[$ and 0 otherwise. We denote by $f(X)$ the classifier mapping X into the $\widehat{Y}$, with $\widehat{Y} = f(X)$. We impose no constraint on the f function: it can be parametric (logistic regression for instance) or not, linear or not, an individual or ensemble classifier which combines a set of homogeneous or heterogeneous models, etc.[8] We only assume that the bank does not use the sensitive attribute D as input for its scoring model.

3.2 Fairness definitions

The literature on designing fair algorithms is extensive and interdisciplinary. Many fairness criteria have been used over the years in computer science, machine learning, criminology, and economics, each aiming to capture different dimensions of fairness (see Hardt et al. (2016), Verma and Rubin

[6] The framework can be easily adapted to a continuous outcome $Y \in \mathcal{Y} \subset \mathbb{R}$, for instance where $Y \in [0, 1]$ denotes the loan interest rate. In this case, conditional probabilities $\Pr(Y = 1|X, D)$ are replaced by conditional expectation $\mathbb{E}(Y = 1|X, D)$.

[7] This setup can be easily extended to a set of q sensitive attributes $D_1, \ldots, D_q$, each of them representing a specific protected group (for instance where D_1 controls for gender, D_2 for age, and so on) or representing the different values of a categorical variable associated to a unique source of potential discrimination (D_1 for Asian-American borrowers, D_2 for African-American borrowers, etc.).

[8] See Lessmann et al. (2015) for an overview and comparison of credit scoring models.

(2018), Kleinberg et al. (2018), Bellamy et al. (2019), Barocas et al. (2020), Lee and Floridi (2020), or Berk et al. (2021)). Here we focus on the formal definitions of four popular fairness metrics which are particularly relevant in credit scoring: statistical parity, equal odds, predictive parity, and overall accuracy.

Definition 1 *A credit scoring model satisfies the* ***statistical parity*** *assumption if the predicted label and the sensitive attribute are independent, i.e., if* $\widehat{Y} \perp\!\!\!\perp D$.

The main idea is that all applicants should have an equivalent opportunity to obtain a good outcome from the credit scoring model, regardless of their group membership. Expressed in terms of conditional probabilities, statistical parity implies:

$$\Pr(\widehat{Y} = 1|D = 1) = \Pr(\widehat{Y} = 1|D = 0) = \Pr(\widehat{Y} = 1) \tag{1}$$

In practice, the difference in opportunity may be due to a composition effect, rather than to discrimination. As a result, it can be more informative to compare similar applicants from protected and unprotected groups, with the comparable individual features (income, job, diploma, etc.). This is precisely the concept of conditional statistical parity.

Definition 2 *A credit scoring model satisfies the* ***conditional statistical parity*** *assumption if the predicted label and the sensitive attribute are independent, controlling for a subset of non-protected attributes* $X_c \subseteq X$, *i.e., if* $\widehat{Y} \perp\!\!\!\perp D|X_c$.

This definition raises the issue of the choice of the conditioning attributes. The goal is to control for the variables that are known to have a first-order impact on creditworthiness in order to constitute homogeneous risk classes. Alternatively, one can rely on a clustering algorithm to partition the individuals into the risk classes, which has the advantage of not selecting the important variables ex-ante. Finally, another approach is to rely on exogenous classification, such as the Basel classification.[9] In all three cases, the statistical parity diagnosis is carried out in each class separately. When all the classes display the same conclusion, the aggregation is trivial. However, when there are some disagreement across classes, an aggregation rule must be used. We can think of two approaches. The first one is to use an economic rule, such as rejecting the hypothesis of global fairness if at least one class displays a significant difference, if the majority of the classes displays

[9] The number of classes considered N plays an important role. Indeed, the larger the N, the more homogenous the sub-groups are (cleaner test), the more likely at least one sub-group is found to be unfair, and the smaller the number of individuals in each sub-group (hence the more difficult it is to find significant results proving lack of fairness).

a significant difference, or if the majority of the individuals belong to classes in which fairness is rejected, etc. A second approach is to rely on a formal statistical test and we will discuss it in the following section.

Statistical parity only focuses on the differences of treatment across groups in terms of output $\widehat{Y}$. The other fairness definitions that we consider below are based on the joint distribution of the triplet $(Y, \widehat{Y}, D)$. By considering the difference between the observed and forecasted outcomes for protected and unprotected groups, these approaches permit to test whether there are some difference of treatment across groups in terms of classification errors.

Definition 3 *A credit scoring model satisfies the* ***equal odds*** *property, if the predicted outcome* $\widehat{Y}$ *and the protected attribute* D *are independent conditional on the actual outcome* Y*, i.e., if* $\widehat{Y} \perp\!\!\!\perp D|Y$.

Unlike statistical parity, equal odds allows $\widehat{Y}$ to depend on D but only through the target variable Y. It implies that clients with a good credit type and clients with a bad credit type should have similar classification, regardless of their (protected or unprotected) group membership. Thus, a credit scoring model is considered fair if the predictor has equal True Positive Rates (TPR, i.e., probability of the truly positive subject to be identified as such) and equal False Positive Rates (FPR, i.e., probability of falsely accepting a negative case).[10] This implies the following two constraints:

$$\Pr(\widehat{Y} = 1|Y = y, D = 0) = \Pr(\widehat{Y} = 1|Y = y, D = 1) = \Pr(\widehat{Y} = 1|Y = y), \quad y \in \{0, 1\} \tag{2}$$

A possible relaxation of equalized odds is to require non-discrimination only within the advantaged outcome group. That is, to require that people who actually pay back their loan, have an equal opportunity of getting the loan in the first place. This relaxation is often called *equal opportunity*. Formally, it implies $\widehat{Y} \perp\!\!\!\perp D\ |Y = 1$ and the equality of the TPRs for the protected and unprotected groups:

$$\Pr(\widehat{Y} = 1|Y = 1, D = 0) = \Pr(\widehat{Y} = 1|Y = 1, D = 1) = \Pr(\widehat{Y} = 1|Y = 1) \tag{3}$$

Conversely, a second relaxation called *predictive equality*, reflects the equality of the FPRs for both groups and corresponds to the assumption $\widehat{Y} \perp\!\!\!\perp D|Y = 0$.

$$\Pr(\widehat{Y} = 1|Y = 0, D = 0) = \Pr(\widehat{Y} = 1|Y = 0, D = 1) = \Pr(\widehat{Y} = 1|Y = 0) \tag{4}$$

[10] The equality of FPR and TPR implies the equality of the odds, i.e., the ratios of probabilities of success (good type) and failure (bad type).

Equal odds, equal opportunity, and predictive equality allow for a perfectly accurate solution $Y = \widehat{Y}$, which is not the case for statistical parity. Thus, it allows aligning fairness with the central goal in supervised learning of building more accurate predictors.[11]

4 Fairness diagnosis

4.1 Fairness inference

If we know the joint distribution of the random variables $(\widehat{Y}, Y, D)$, we can determine without any ambiguity whether this joint distribution satisfies any fairness definition. However, in practice we must take this decision from a sample. Then, the implementation of the previous fairness definitions requires being able to conclude from this sample whether the assumptions of (conditional) independence between the variables $\widehat{Y}$, Y, and D are satisfied or not. Here, we propose a general testing methodology that considers estimation uncertainty to statistically test for the fairness of a credit scoring model.

We consider a sample S_n of n observations $\{y_j, x_j, d_j\}_{j=1}^n$ issued from the joint distribution $p_{Y,X,D}$, where the index j denotes the j^{th} credit applicant. Based on this sample, the credit scoring model produces a set of decisions $\{\widehat{y}_j\}_{j=1}^n$. Considering the sample $\{\widehat{y}_j, y_j, d_j\}_{j=1}^n$, we wish to test whether the credit scoring model satisfies a particular fairness definition indexed by $i \in \{SP, CSP, EO, EOP, PE\}$ with:

$$\mathrm{H}_{0,SP} : \widehat{Y} \perp\!\!\!\perp D \qquad \mathrm{H}_{0,CSP} : \widehat{Y} \perp\!\!\!\perp D|X_c \qquad \mathrm{H}_{0,EO} : \widehat{Y} \perp\!\!\!\perp D|Y$$

$$\mathrm{H}_{0,EOP} : \widehat{Y} \perp\!\!\!\perp D|Y = 1 \qquad \mathrm{H}_{0,PE} : \widehat{Y} \perp\!\!\!\perp D|Y = 0$$

where SP stands for statistical parity, CSP for conditional statistical parity, EO for equal odds, EOP for equal opportunity, and PE for predictive equality. Formally, we denote a fairness test statistic as:

$$F_{H_{0,i}} \equiv h_i(\widehat{Y}_j, Y_j, D_j; j = 1, \ldots, n) = h_i\left(f\left(X_j\right), Y_j, D_j; j = 1, \ldots, n\right) \tag{5}$$

where $h_i\left(.\right)$ denotes a functional form that depends on the null hypothesis $\mathrm{H}_{0,i}$ which is considered, the scoring model f, and the sample $\{\widehat{y}_j, y_j, d_j\}_{j=1}^n$. As all fairness metrics can be expressed in terms of independence assumptions, we can derive specific inference. Under the null hypothesis $\mathrm{H}_{0,i}$, we

[11] The impossibility theorem of Kleinberg et al. (2017) states that no more than one of the three fairness definitions of statistical parity, predictive parity, and equal odds can hold at the same time for a well calibrated classifier and a given protected attribute.

assume that the test statistic $F_{H_{0,i}}$ has a $\mathcal{F}_i$ distribution, and we denote $d_{1-\alpha}$ the corresponding critical value at $\alpha\%$ significance level.

To the best of our knowledge, this is the first time that a formal inference procedure is introduced in the context of fairness assessment. Such approach offers several advantages. First, whatever the test statistic considered, fairness inference allows us to consider estimation uncertainty when comparing conditional probabilities for protected and unprotected groups. Second, it allows to fix the probability of incorrectly deciding that a scoring model is unfair. Third, beyond estimation risk, the need for inference in fairness evaluation comes from the fact that most metrics are conditional. Thus, their assessment implies a comparison of outcomes for different classes of the credit applicants, where each class corresponding to a particular set of values of the conditioning variable(s). For instance, conditional statistical parity implies to test the independence between the predicted default $\widehat{Y}$ and the protected attribute D for different classes of credit applicants, pooled according to the borrower and loans characteristics. In this context, fairness test statistics are useful because they allow for a global (aggregate) diagnosis without regard to utility functions.

The notation for F_i encompasses a wide class of test statistics which can be implemented in this context. For instance, one can use the chi-squared conditional independence test, the Cochran–Mantel–Haenszel (CMH) test, the z-tests associated to the hypothesis tests of the difference, ratio, or odds ratio of two independent proportions, or a likelihood ratio test. For ease of presentation, in the sequel we only consider standard chi-squared conditional independence tests (cf. Appendix A.1 for a formal presentation).

As an illustration, let us a consider a numerical example. We assume that a regulator whishes to test whether a scoring model satisfies the null hypothesis of conditional statistical parity $\mathrm{H}_{0,CSP}$ by considering a sample $1,000$ individuals, among which 310 belong to the protected group. These individuals are divided into two classes, labeled C_1 (827 borrowers) and C_2 (173 borrowers) respectively, according to their characteristics X_c. The empirical distribution of the predicted oucome $\hat{Y}$ and the protected attribute D is displayed in Figure (1) for the two groups.

Insert Figure 1

For each class, we compute the standard Pearson's chi-squared test statistics, respectiveley denoted χ_1^2 and χ_2^2, by comparing the empirical distribution of $(\hat{Y}, D)$ to the theoretical distribution under the null of independence. We get a realization of $\chi_1^2 = 13.15$ for the first class and $\chi_2^2 = 3.24$

for the second one.[12] Under the null, each individual statistic follows a $\chi^2_{(1)}$ distribution. For a 5% significance level, the critical value is equal to $d^2_{1,0.05} = 3.84$, thus we reject the null hypothesis $\widehat{Y} \perp\!\!\!\perp D$ for the borrowers in class C_1, whereas we do not reject the null for the borrowers in class C_2. To aggregate these diagnoses, we compute a conditional test statistic $F_{H_{0,CSP}}$ simply defined as the sum of the individual chi-squared statistics, i.e., $F_{H_{0,CSP}} = \chi^2_1 + \chi^2_2 \sim \chi^2_{(2)}$. As $F_{H_{0,CSP}} = 16.39 > d^2_{2,0.05} = 5.99$, we reject the null hypothesis $\mathrm{H}_{0,CSP}$ of conditional statistical parity, i.e., $\widehat{Y} \perp\!\!\!\perp D|X_c$. This example illustrates the fact that the choice of the conditional test statistics implies a particular ordering (or preference relationship) on the cross-sectional profile of the fairness diagnostic among classes.

4.2 Fairness interpretability

Interpretability is a the heart of the financial regulators' current concerns about the governance of AI, especially in the credit scoring industry. Here, we define interpretability as the degree to which a human can understand the cause of a decision (Miller (2019)). It is important to note that interpretability does not necessarily implies fairness. Even if the decisions issued from a model are interpretable, they may treat unfavorably a group of users sharing a protected attribute. Conversely, the absence of discrimination is compatible with both interpretable or non-interpretable models.

There are different ways to interpret machine learning models and their decisions. We can distinguish between interpretable models and uninterpreatable models. Intrinsic interpretability refers to ML models that are considered interpretable due to their simple structure, such as decision trees or linear models. Differently, black-box models, such as random forests or deep neural networks, have observable input-output relationships, but lack clarity around inner workings. This means that clients, regulators, and other stakeholders, even those who design the models, cannot easily understand how variables are being combined to make the risk predictions and to deliver the credit approval decision. To allow interpreting predictions of black box ML models, various model-agnostic methods have been developed (see Molnar (2019) for an overview). For instance, the *Partial Dependence Plot* (PDP) displays the marginal impact of a specific feature on the outcome $\widehat{Y}$ of a model.

[12] Among the 827 borrowers in C_1, 611 have been classified as low-risk ($\widehat{Y} = 1$) by the model, 216 as high-risk ($\widehat{Y} = 0$), and 270 borrowers belong to the protected group. The standard Pearson's chi-squared test statistic (see Appendix A.1 for more details) for the class C_1 is:

$$\begin{aligned}\chi^2_1 &= (433 - 557 \times 611/827)^2/(557 \times 611/827) + (124 - 557 \times 216/827)^2/(557 \times 216/827) \\ &\quad +(178 - 611 \times 270/827)^2/(611 \times 270/827) + (92 - 270 \times 216/827)^2/(270 \times 216/827) = 13.15\end{aligned}$$

This plot is useful to explore whether the relationship between the target and a feature is linear or more complex. These methods have the advantage to be model-free in the sense that they allow explaining predictions of arbitrary ML models independently of its form, its implementation, or its internal model parameters.

Instead of focusing on the interpretability of the outcome of the model, we focus on the interpretability of its fairness metric. We define fairness interpretability as the degree to which an applicant, a regulator, or any stakeholder, can understand the cause of a discrimination with respect to a given protect attribute induced by a machine learning model, if there exists. Here, we propose a simple model-agnostic method which we call *Fairness Partial Dependence Plot* (FPDP). The latter displays the marginal effect of a specific feature on a fairness diagnosis test associated to a credit scoring model. The logic of the FPDP is similar to that of PDP, except that we explore the relationship between one feature and a fairness test output. The goal of the FPDP is to identify the feature(s) at the origin of the discrimination bias and the rejection of the null hypothesis of fairness, whatever the hypothesis considered. These variables are called candidate features.

Formally, denote by X_A the feature for which we want to measure the marginal effect, X_B are the other features, and $f(.)$ is the credit scoring model such that $\widehat{Y} = f(X_A, X_B)$. Rewrite the fairness test statistic as:

$$\begin{aligned} F_{H_{0,i}} &\equiv h_i(\widehat{Y}_j, Y_j, D_j; j = 1, \ldots, n) && (6) \\ &= h_i(f(X_{A,j}, X_{B,j}), Y_j, D_j; j = 1, \ldots, n) && (7) \\ &= \widetilde{h}_i(X_{A,j}, X_{B,j}, Y_j, D_j; j = 1, \ldots, n) && (8) \end{aligned}$$

with $\widetilde{h}(.)$ a nonlinear positive function. We have to distinguish two cases depending on whether X_A is a categorical feature or a continuous feature.

Definition 4 *Consider a categorical feature $X_A \in \{c_1, \ldots, c_p\}$, the corresponding FPDP is:*

$$F_{H_{0,i}}(c_k) = \widetilde{h}_i((c_k, x_{B,1}, y_1, d_1), \ldots, (c_k, x_{B,n}, y_n, d_n)) \quad \forall k = 1, \ldots, p \tag{9}$$

Consider a continuous feature $X_A \in \left[x_A^{\min}, x_A^{\max}\right]$, the corresponding FPDP is defined as:

$$F_{H_{0,i}}(x) = \widetilde{h}_i((x, x_{B,1}, y_1, d_1), \ldots, (x, x_{B,n}, y_n, d_n)) \quad \forall x \in \left[x_A^{\min}, x_A^{\max}\right] \tag{10}$$

The FPDP associated to the c_k category displays the realization of the fairness test statistic obtained when the categorical feature X_A takes a value c_k for all instances, whatever their other

features. In the continuous case, the FPDP associated to the value x corresponds to the fairness test statistic obtained when the feature X_A takes the value x for all instances, whatever their other features. Notice that for a given scoring model, the FPDP depends on the null hypothesis i which is considered. Thus, we can define a FPDP with respect to (conditional) statistical parity, equal odds, equal opportunity, or predictive equality.

The FPDP allows to identify the feature(s), called candidate variable(s), which are causing the lack of fairness. Whatever the type of the feature (continuous or categorical), consider a case for which the null of fairness is initialy rejected as $F_{H_{0,i}} > d_{1-\alpha}$ where $d_{1-\alpha}$ is the critical value of the test for $\alpha\%$ significance level associated to the null distribution $\mathcal{F}_i$. A given feature is identified as a candidate variable, if changing its value reverses the fairness diagnosis.

Definition 5 *A feature X_A is considered as a candidate variable, if there exists at least one value $c_k \in \{c_1, \ldots, c_p\}$ (categorical variable) or $x \in \left[x_A^{\min}, x_A^{\max}\right]$ (continuous variable), such that $F_{H_{0,i}}(c_k) < d_{1-\alpha}$ or $F_{H_{0,i}}(x) < d_{1-\alpha}$.*

The FPDP has two main advantages. First, FPDP is a model agnostic method. Regardless of the ML algorithm used by data scientists, the fairness of the credit approval decision can always be assessed. Second, FPDP breaks correlations among features. Indeed, in case of disparate impact, some features are correlated with the protected attribute and leads to significantly different outcomes for protected and unprotected groups. Disparate impact can arise in two cases: either when one feature correlates with the protected attribute, or when a set of features collectively proxy for the protected attribute. The latter is likely to occur in flexible, non-linear ML algorithms. Thus, breaking the correlations between the features and the protected attribute, and/or among these proxy features allows us to identify the candidate variables, as it induces a change in the fairness diagnosis. Thus, contrary to standard PDP, FPDP does not assume that the features are independent.[13]

[13] In a standard PDP analysis, it is assumed that features are independent. This assumption is problematic as such analysis overlooks the indirect effects of a feature (through other features) on the outcome.

5 Application

5.1 Data

We illustrate our methodology using the German Credit Dataset (Dua and Graff (2019)), which includes 1,000 consumer loans extended to respectively 310 women and 690 men.[14] For each applicant, we know his or her actual credit risk type, called *credit risk*: good-type or low-risk ($Y = 1$) vs. bad-type or high-risk ($Y = 0$) (see Verma and Rubin (2018) for details). This variable is our target variable. In total, 300 borrowers are in default ($Y = 0$) among which 191 are men and 109 women. Moreover, there are 19 explanatory variables measuring either some attributes of the borrower (e.g., gender, age, occupation, credit history) or some characteristics of the loan contract (e.g., amount, duration). More information about the database can be found in Table A.2 in the appendix.

Insert Figures 2 and 3

We start by contrasting in Figure *2* the distributions of each of the 20 variables for men and women. We clearly see that the default rate is higher for women (35.16%) than for men (27.68%). Moreover, women borrowers tend to be younger and to exhibit a lower home-ownership rate and employment duration. However, being correlated with gender does not imply that this feature necessarily leads to a different treatment between men and women. It is only when this feature is also driving the target variable that it may lead to discrimination. As a result, one needs to go beyond histograms and quantify the association with default.

In Figure *3*, we present a scatter plot of Cramer's V, which is a measure of association between two variables.[15] The X-axis displays the association between each explanatory variable and the target variable whereas the Y-axis does so for each explanatory variable and gender. By construction, the top-right region of the plot includes variables that are both gendered and with a strong ability to classify between good and bad-type borrowers. While such variables would be natural candidates for inducing a classification algorithm to put any woman borrower in the high-risk category, we have no obvious candidate example. Moreover, central to this analysis is the substitution that exists between being associated with gender or with default. Hence, one may think about these variables in terms of "indifference curves" connecting variables displaying a comparable level of

[14] In the initial database, the gender and the marital status of the applicants are specified in a common attribute with five categorical values (single male, married male, divorced male, single female, married or divorced female). Here, as we focus on gender discrimination whatever the marital status, we consider a binary variable representing the single, married, or divorced females (protected group) versus the single, married, or divorced males (unprotected group). The original database can be found here.

[15] The Cramer's V varies from 0 (corresponding to no association between the variables) to 1 (complete association).

potential impact on fairness: either being strongly associated to gender but mildly to default (e.g., age), being moderately associated with both (e.g., purpose), or being mildly associated to gender but strongly to default (e.g., account status). Everything else held constant, variables located on an indifference curve further away from the origin are more likely to lead the algorithm to grant a less favorable credit score to women than to men.

5.2 Credit scoring models

We place ourselves in the configuration of a bank that seeks to assess the creditworthiness of its customers through the development of a credit scoring model. Before modelling credit default, we apply various pre-treatments to the dataset. First, we transform the 11 categorical variables into binary variables. Doing so is standard practice as it gives more degrees of freedom to the algorithms and permits to better exploit their inherent flexibility. Second, we remove the binary variable foreign worker from the set of features, as we do not want to mix two possible sources of discrimination. Such pre-treatment leads us with a total of 55 explanatory variables. In general, the dataset is split into a training subsamble (in-sample calibration) and a testing subsamble (out-of-sample estimation). Here, we deviate from this and estimate all scoring models using the whole sample in order to have enough observations when conducting conditional fairness tests.

We estimate a set of increasingly sophisticated credit scoring models to predict default (see Lessmann et al. (2015) for a survey), namely logistic regression (LR), penalized logistic regression (Ridge), classification tree (TREE), random forests (RF), XGBoost (XGB), support vector machine (SVM), and artificial neural networks (ANN). Thus, we consider both standard parametric regression models (LR and Ridge) and machine-learning models, able to extract non-linear relationships. We estimate individual classifiers (SVM, TREE, ANN) as well as ensemble methods (RF, XGB), which have proved to perform well for credit scoring applications (Lessmann et al. (2015)). Most importantly, our analysis includes intrinsically interpretable models or white-box models (LR, Ridge, TREE) and black-box models (RF, ANN) as our fairness diagnosis method can accomodate both types of models. In practice, the performance of ML models (and, as shown below, their fairness diagnosis) is quite sensitive to the value of the parameters used to fine tune the model and control the learning process. In the present study, we determine hyperparameter values using a ten-fold cross-validation and a random search algorithm.[16] For ANN, we standardize continuous features to

[16] We split the dataset into ten subsamples. We use nine of them for in-sample calibration, while using the remaining one for out-of-sample testing. This procedure is carried out ten times by changing the subsample used out-of-sample.

speed up the convergence of the optimization algorithm.

We first estimate the credit scoring models by including *gender* in the list of the explanatory variables and refer to these models as the *with-models*. While this may not be a particularly realistic setting, as banking regulatory authorities typically prevent lenders from including gender in scoring models, it constitutes a useful reference point in our controlled experiment. As these models will ultimately be used to check whether our fairness tests are able to detect disparate treatment, we make sure that the gender variable is selected in each model. As a result, we may not select the best performing models among all possible ones. Panel A of Table 1 displays the performance of the seven considered with-models. we measure performance using the percentage of correct classification (PCC) and the area under the curve (AUC). We see that all considered models perform well as the PCC ranges between 76.4 and 87.3 and the AUC between 0.811 and 0.938. The best performance is achieved by the RF model and the lowest one by the Ridge logistic regression model.

In a second step, we rerun all models after removing gender from the analysis, and we call these models the *without-models*. In this case, if a gender discrimination occurs, it is necessarily through indirect discrimination due to features being correlated with the protected attribute. Panel B of Table 1 displays the PCC and AUC values for the seven without-models. Overall, the performance remains quite good, as none of the performance measures changes by more than five percentage points and RF remains the best performing model.[17]

Insert Table 1

5.3 Fairness diagnosis

We now turn to implementing our inference tests for algorithmic fairness derived in Section 4.1. To do so, we use the scoring models estimated in the previous section as test algorithms. We first verify whether there is disparate treatment by considering the six alternative fairness measures described in Section 3.1 (i.e., statistical parity, conditional parity, equal odds, predictive equality, and equal opportunity) and the seven with-models.

We report in Table 2 for each with-model the p-value associated with a given null hypothesis of fairness. Overall, the message is clear as the null hypothesis of fairness is rejected at the 95%

Within this process, the random search algorithm trains the considered credit scoring models based on different hyperparameter settings. Finally, the random search algorithm chooses the hyperparameter values with the highest average accuracy across all subsamples. See Appendix A.5 for more information about hyperparameters.

[17] As we only consider with-models that selects the gender variable, we disregard some models with a higher in-sample performance. We do not have similar constraints when dealing with without-models. As a result, some models have a better in-sample performance without gender than with gender.

confidence level for most model - fairness measures combinations. These results confirm the good performance of our tests whatever the scoring model and the fairness definition considered. One exception is when we split the sample into two groups using a K-Prototypes clustering algorithm in order to test for conditional statistical parity. Group 2 gathers individuals borrowing relatively higher amounts and over longer periods and exhibit a more unstable credit history (cf. Appendix A.3). The result of the conditional statistical parity test indicates that we cannot reject the fairness null hypothesis in Group 2. This is due to the fact that most of these individuals are so high-risk that their gender no longer influences the model predictions. Conversely, individuals in Group 1 exhibit more diverse risk profiles and display significant rejection rates between men and women. At the aggregate level, the global conditional parity test leads to reject the null hypothesis of fairness for all models at the 95% confidence level.

Insert Table 2

To know whether the scoring models also leads to indirect discrimination, we remove *gender* from the scoring models. Table 3 displays similar results as in Table 2 but for *without*-models. For most models, we do not reject anymore the null hypothesis of statistical parity. This suggests that the lack of fairness previously detected was actually due to direct discrimination. We reach a similar conclusion using conditional statistical parity, equal odds, equal opportunity, and predictive equality.[18] Differently, with the ANN model, we reject the null fairness hypothesis as the p-values associated with (conditional) statistical parity are around 0.01. For this model, removing the gender variable is not a sufficient condition to safeguard members of the protected group. There is indeed evidence for indirect discrimination, most likely through the combination of variables able to replicate or produce a synthetic version of the gender variable. The fact that the ANN is the only model for which gender discrimination occurs is not a coincidence, as this model is one of the most flexible classifiers considered in the present study.

Insert Tables 3 *and* 4

In a final step, we illustrate the important role played by operational risk. Indeed, we show that the choice of the hyperparameters used to control the learning process of credit scoring models

[18] This result may be surprising at first glance given the fact that the impossibility theorem of Kleinberg et al. (2017) states that no more than one fairness definition can hold at the same time. However, the impossibility theorem concerns the true joint distribution of $(Y, \widehat{Y}, D)$ that we do not observe in practice. When using a finite sample of loans and accounting for estimation risk, it is possible to not reject the null for two or more fairness definitions. Inference here allows us to consider the uncertainty associated with the estimates, while controlling for the risk of falsely rejecting the fairness null hypothesis.

can have strong consequences in terms of fairness. To illustrate our point, we consider an alternative TREE model, call TREE prime, which is based on a slightly modified set of hyperparameters. Specifically, TREE prime relies on a procedure commonly used in the credit scoring industry (Lessmann et al. (2015)), based on a k-fold cross-validation and a random search algorithm.[19] This alternative technique (1) is state-of-the-art and could be put in production by a lender, (2) is in some dimensions more general , and in other dimensions more restrictive than the original one, (3) leads to a slightly lower, yet quite good, performance, PCC = 79.0 vs. 81.5. However, we see in Table *4* that it leads to drastically different conclusions in terms of fairness. Indeed, for TREE prime, we reject the null hypothesis of fairness for virtually all metrics at the 95% confidence level.

5.4 Interpretability and mitigation

In the previous section, we have identified several scoring models that were classified as unfair by our testing procedure. We are now going to identify the variable(s) that are at the origin of the fairness concern. As an example, we consider the TREE-prime model as it has been shown to exhibit a significant lack of fairness. This conclusion was reached for statistical parity, conditional statistical parity, equal odds, and equal opportunity. As in any decision tree, the underlying process of construction of the model as a recursive partitioning of the space of explanatory variables can be plotted as a tree diagram (See Figure *7* in the Appendix). We observe that 14 features have been selected by the TREE algorithm, namely housing, number of credits, credit amount, telephone, installment rate, property, age, other installment plans, account status, credit duration, credit history, purpose, saving, and employment duration.

We now implement the FPDP associated to the statistical parity null hypothesis. The individual plots associated with the 14 considered features are reported in Figure *4*. Recall that in the initial sample (see Table *4*), the statistical parity test leads to the rejection of the null hypothesis of fairness or in other words, the p-value is smaller than the 10% threshold. As explained in the previous section, a feature is said to be a candidate variable if setting the same value of this variable to all borrowers leads to the non-rejection of the null hypothesis (i.e., p-value $>$ 10%).

As shown in Figure *4*, we identify six candidate variables, namely credit duration, credit history, purpose, savings, account status, and telephone.[20] The analysis of the decision tree reported in the

[19] We reduce the set of values for the maximum depth of the tree from 1-29 to 1-9, we increase the set of values for the minimum number of instances required to split a node from 2-9 to 2-59, and we increase the set of values for the minimum number of individuals by leaf from 1-19 to 1-59.

[20] For the account-status variable, the test statistic cannot be defined for one modality since all predicted outcomes are equal to 1, meaning that it is independent from gender (no rejection of the null hypothesis).

Appendix confirms the consistency of the results issued from FPDPs. First, features not identified as candidate variable (e.g. credit amount) are only included in paths where all leaf nodes are associated to the same class label ($\widehat{Y} = 0$ or $\widehat{Y} = 1$). Second, all the candidate variables are features which partition the data space into leaves associated to positive and negative labels. Third, all the features used to split the credits between positive and negative lables are not necessarily identified as candidate variables (e.g., property). In our case, only credit history is identified as candidate variable, which illustrates the added value of our FPDP analysis.[21]

Insert Figures 4 and 5

In a final stage, we show that the variables identified as causing the statistically significant difference in rejection rates between men and women are not all consistent with business necessity. As a result, our method indicates that this particular algorithm is, in this sample, violating the fair-lending principle and can be qualified as illegal. To prove this, we start by highlighting in red, in Figure 5, the six features identified as candidate variables. We see that five of them are significantly correlated with the target variable and with gender. These variables correspond to legitimate variables which are theoretically related to default and regularly used in credit scoring models. In other words, they constitute legitimate variables when forecasting the default of a pool of retail borrowers. Differently, the last candidate variable, telephone, appears to be much less correlated with the target and gender variables. Unlike the other five candidate variables, the causal effect of having a telephone and defaulting on a loan is not completely straightforward. What is more likely though is the fact that both phone and default exhibit a common cause, which explains why telephone has been used by the algorithm to split the data space into leaves associated to positive and negative labels.

In an attempt to mitigate the fairness problem, we remove the telephone variable from the decision tree. It is important to notice that we do not re-estimate the model but we simply exclude this particular variable when partitioning the data space. Doing so leads to a new set of estimated labels ($\widehat{Y}$), on which we run our inference tests. We see in Table *4* that removing telephone has an important effect in terms of fairness as we cannot reject anymore the null hypothesis of fairness for

[21] As a robustness check, we carry out similar FPDP analyses for the fairness definitions in the Appendix. In Figure A.6, we report the FPDP associated to the null hypothesis of conditional statistical parity, equal odds, and equal opportunity. The main takeaway is that we identify the same six candidate variables as those identified for statistical parity, whatever the fairness definition considered. Since these variables are both correlated with gender and default, they induce gender discrimination not only in the predicted default probabilities, but also in the FPR, TPR, and accuracy rates.

most measures. On the other hand, the effect on performance is trivial.

6 Conclusion

Credit scoring algorithms can be life-changing for many households and businesses. Indeed, they impose conditions on who can access credit and at which terms. As a result, it is of primary importance to make sure that algorithms comply with the fair lending principles written in the law, and even so when they are based on complex and opaque ML techniques and big data. In this paper, we propose a framework to formally check whether there exists a statistical significant difference in terms of rejection rates or interest rates between protected and unprotected groups and whether this difference is only due to credit worthiness. Our framework provides guidance on how algorithmic fairness can be monitored by lenders, controlled by their regulators, and improved for the benefit of protected groups.

The high legal and regulatory uncertainties surrounding the use of ML algorithms acts as an impediment for financial service providers to innovate and invest in screening technologies (Evans (2017), Bartlett et al. (2021b)). Furthermore, recent ruling in Europe against companies using discriminatory algorithms (Geiger (2021)) as well as debate in the New York City Council to potentially ban some automated tools used by corporations (Givens et al. (2021)) put a spotlight on this concerns. As algorithmic discrimination of women and minorities may be completely unintentional and can be embedded in data and/or in the inner workings of algorithms, it is more important than ever for lenders and their regulators to benefit from clear guidelines and tools able to redflag any form of non-legitimate discrimination. We believe our methodology can contribute to provide such much needed guidelines and regulatory tools.

While we focus on access to credit, this methodology also can prove useful for studying other life-changing decision-making algorithms. Indeed, similar algorithms are in use in the fields of predictive justice (sentencing and probation), hiring (automated pre-screening of applicants), education (university admission and scholarship granting), and housing (tenant selection).

References

ACPR (2020). Governance of artificial intelligence in finance. Discussion papers publication, November, 2020.

Alesina, A. F., Lotti, F., and Mistrulli, P. E. (2013). Do Women Pay More for Credit? Evidence From Italy. *Journal of the European Economic Association*, 11(s1):45–66.

Arrow, K. (1971). The Theory of Discrimination. *Applied Economics*, 6.

Barocas, S., Hardt, M., and Narayanan, A. (2020). *Fairness and Machine Learning*. `http://www.fairmlbook.org`.

Bartlett, R., Morse, A., Stanton, R., and Wallace, N. (2021a). Consumer-lending discrimination in the fintech era. *Journal of Financial Economics, forthcoming*.

Bartlett, R. P., Morse, A., Wallace, N., and Stanton, R. (2021b). Algorithmic accountability: A legal and economic framework. *Berkeley Technology Law Journal, forthcoming*.

Bayer, P., Ferreira, F., and Ross, S. L. (2018). What Drives Racial and Ethnic Differences in High-Cost Mortgages? The Role of High-Risk Lenders. *Review of Financial Studies*, 31(1):175–205.

Becker, G. S. (1957). *The Economics of Discrimination*. Chicago: The University of Chicago Press.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):1–15.

Berg, T., Burg, V., Gombović, A., and Puri, M. (2020a). On the Rise of FinTechs: Credit Scoring Using Digital Footprints. *Review of Financial Studies*, 33(7):2845–2897.

Berg, T., Puri, M., and Rocholl, J. (2020b). Loan Officer Incentives, Internal Rating Models, and Default Rates. *Review of Finance*, 24(3):529–578.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2021). Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50(1):3–44. Publisher: SAGE Publications Inc.

Bhutta, N. and Hizmo, A. (2021). Do minorities pay more for mortgages? *Review of Financial Studies*, 34(2):763–789.

Cowgill, B. and Tucker, C. E. (2020). Algorithmic Fairness and Economics. *SSRN Electronic Journal*.

D'Acunto, F., Ghosh, P., Jain, R., and Rossi, A. G. (2021). How Costly are Cultural Biases? *SSRN Electronic Journal*.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*.

Deloitte (2020). Building trustworthy AI. `https://www2.deloitte.com/fr/fr/pages/services-financier/articles/building-trustworthy-ai.html`.

Dobbie, W., Liberman, A., Paravisini, D., and Pathania, V. (2018). Measuring Bias in Consumer Lending. Working Paper 24953, National Bureau of Economic Research.

Dua, D. and Graff, C. (2019). UCI machine learning repository. `http://archive.ics.uci.edu/ml`.

EBA (2020). Report on big data and advanced analytics. European Banking Authority, January, 2020.

EC (2020). White paper on artificial intelligence: A european approach to excellence and trust. European Commission, February, 2020.

EEOC (1978). Uniform guidelines on employee selection procedure. Equal Employment Opportunity Commission, August, 1978.

Evans, C. A. (2017). Keeping Fintech Fair: Thinking About Fair Lending and UDAP Risks. Consumer Compliance Outlook, Federal Reserve System.

Evans, C. A. and Miller, W. (2019). From Catalogs to Clicks: The Fair Lending Implications of Targeted, Internet Marketing. Consumer Compliance Outlook, Federal Reserve System.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2021). Predictably Unequal? The Effects of Machine Learning on Credit Markets. *Journal of Finance, forthcoming*.

Fuster, A., Plosser, M., Schnabl, P., and Vickery, J. (2019). The Role of Technology in Mortgage Lending. *Review of Financial Studies*, 32(5):1854–1899.

Geiger, G. (2021). Court Rules Deliveroo Used 'Discriminatory' Algorithm. https://www.vice.com/.

Gillis, T. B. and Spiess, J. (2018). Big Data and Discrimination. *SSRN Electronic Journal*.

Givens, A. R., Schellmann, H., and Stoyanovich, J. (2021). Opinion | We Need Laws to Take On Racism and Sexism in Hiring Technology. *The New York Times*.

Hardt, M., Price, E., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323. Curran Associates, Inc.

Jansen, M., Nguyen, H., and Shams, A. (2020). Rise of the Machines: The Impact of Automated Underwriting. Working paper, Ohio State University.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Rambachan, A. (2018). Algorithmic Fairness. *AEA Papers and Proceedings*, 108:22–27.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *ArXiv*, abs/1609.05807.

Lael, B. (2021). Supporting responsible use of ai and equitable outcomes in financial services. Board of Governors of the Federal Reserve System.

Lee, M. and Floridi, L. (2020). Algorithmic fairness in mortgage lending: From absolute conditions to relational trade-offs. *SSRN Electronic Journal*.

Lessmann, S., Baesens, B., Seow, H.-V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research*, 247(1):124 – 136.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1 – 38.

Molnar, C. (2019). *Interpretable Machine Learning*. `https://christophm.github.io/interpretable-ml-book/`.

Morse, A. and Pence, K. (2020). Technological innovation and discrimination in household finance. Working Paper 26739, National Bureau of Economic Research.

O'Neil, C. (2017). *Weapons of Maths Destruction: How Big Data Increases Inequality and Threatens Democracy*. Wiley.

Phelps, E. (1972). The Statistical Theory of Racism and Sexism. *American Economic Review*, 62:659–61.

Philippon, T. (2019). On fintech and financial inclusion. Working Paper 26330, National Bureau of Economic Research.

Prince, A. and Schwarcz, D. (2020). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, 105:1257–1301.

PWC (2021). Model risk management of AI and machine learning systems. `https://www.pwc.co.uk/issues/data-analytics/insights/model-risk-management-of-ai-and-machine-learning-systems.html`.

Ravina, E. (2019). Love & Loans: The Effect of Beauty and Personal Characteristics in Credit Markets. *SSRN Electronic Journal*.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7.

Vigdor, N. (2020). Apple card investigated after gender discrimination complaints. *New York Times*.

Figure 1: Numerical example of joint distribution between predicted outcome and protected attribute for two classes of borrowers

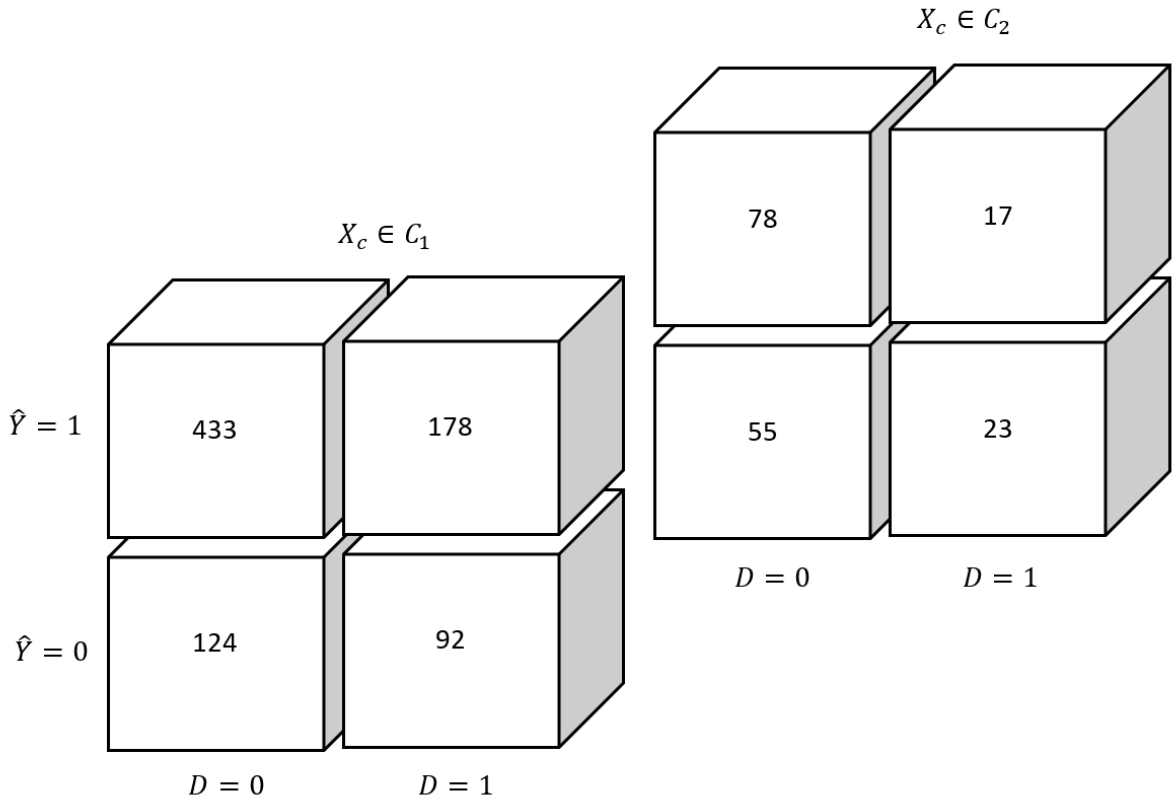
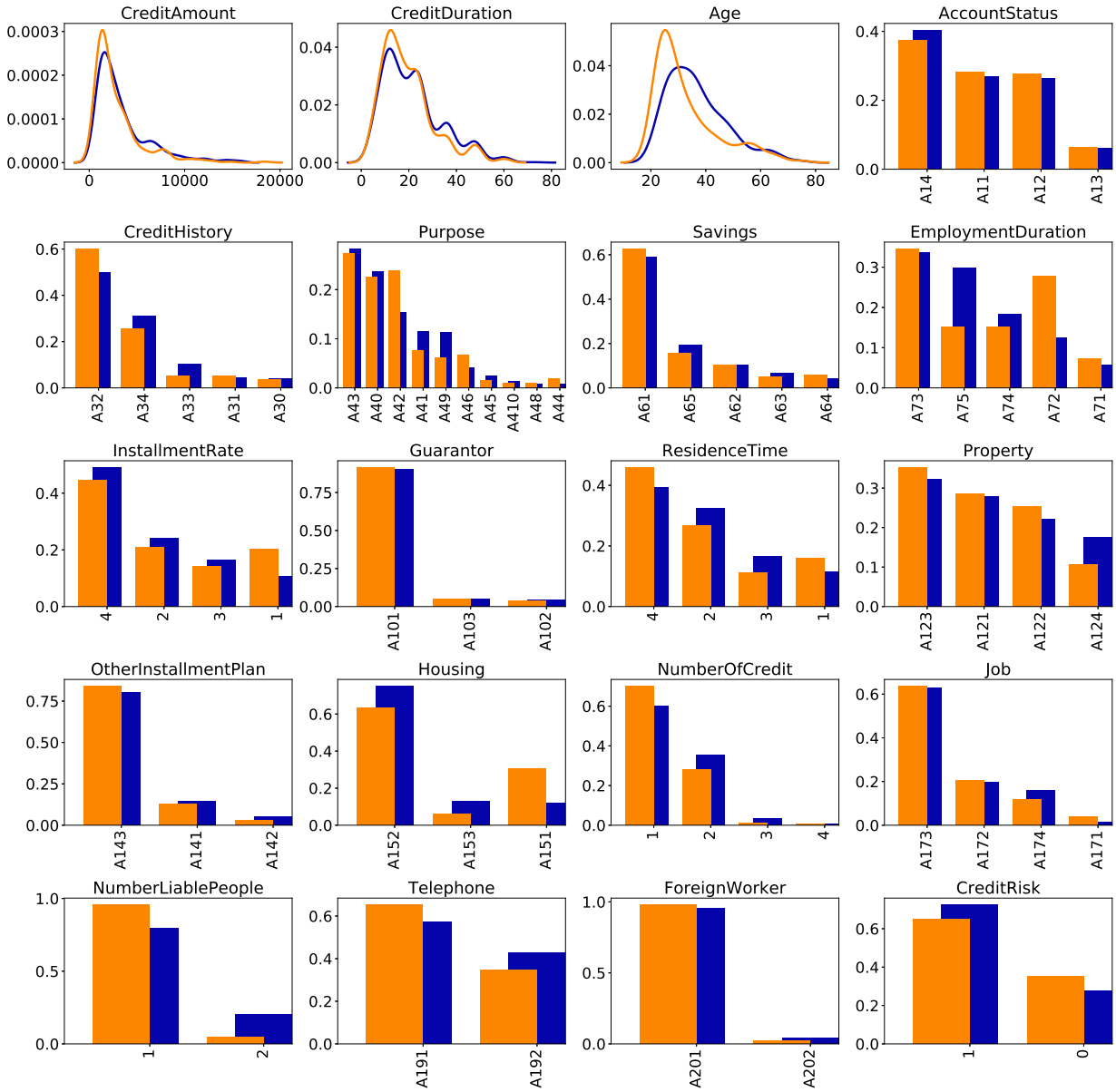
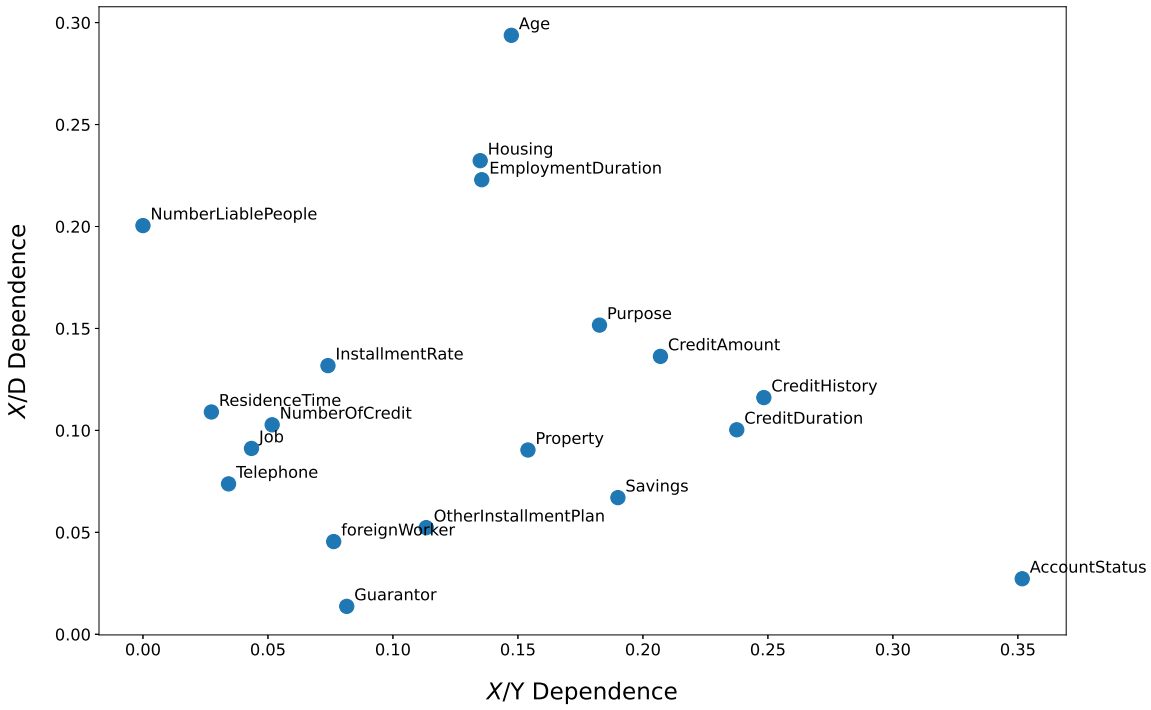


Figure 2: Feature Distributions



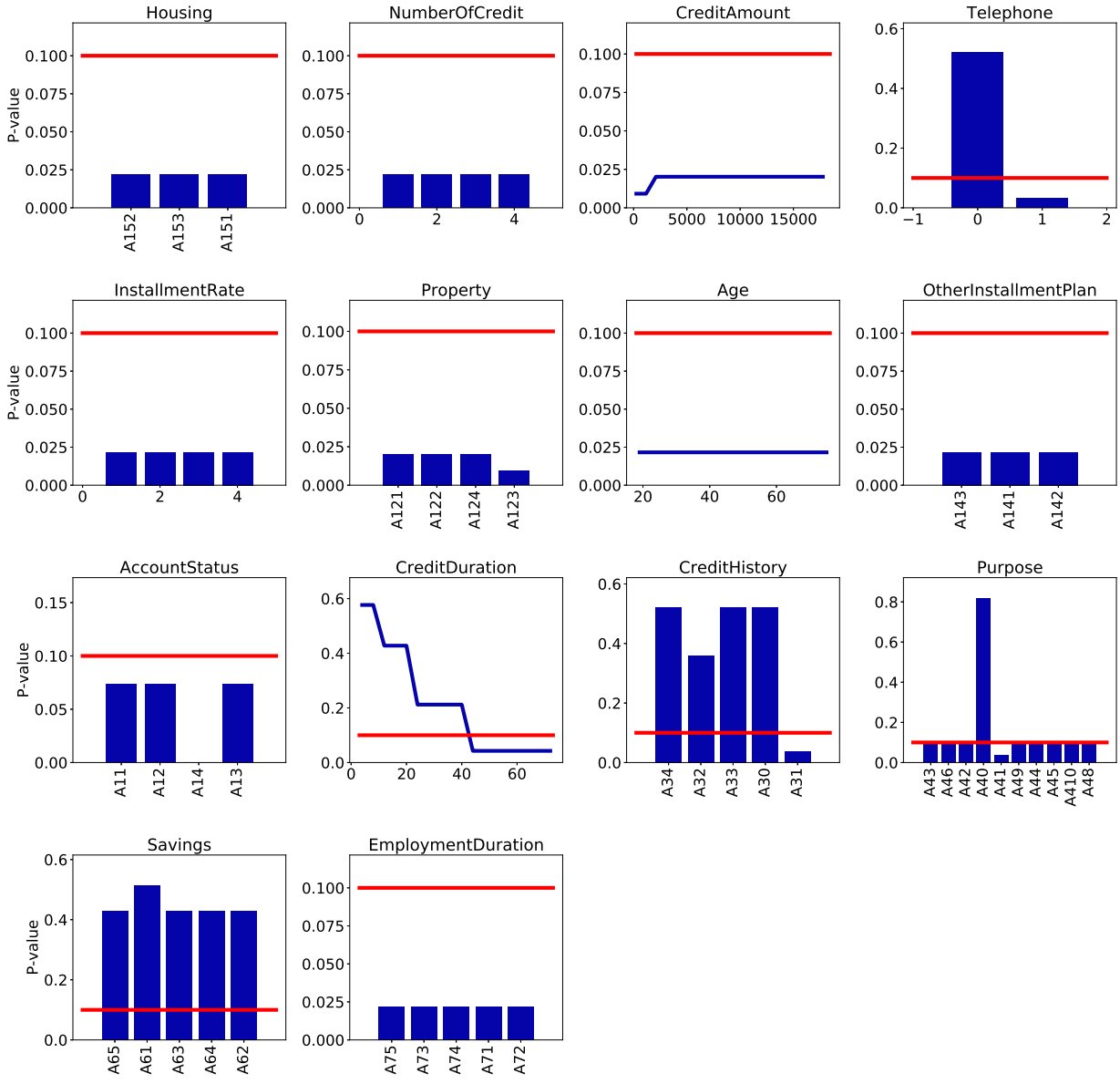
Notes: These figures display the feature distributions by gender, using kernel density estimation for continuous variables. Blue color refers to men and orange to women.

Figure 3: Measures of association between features, target variables, and gender



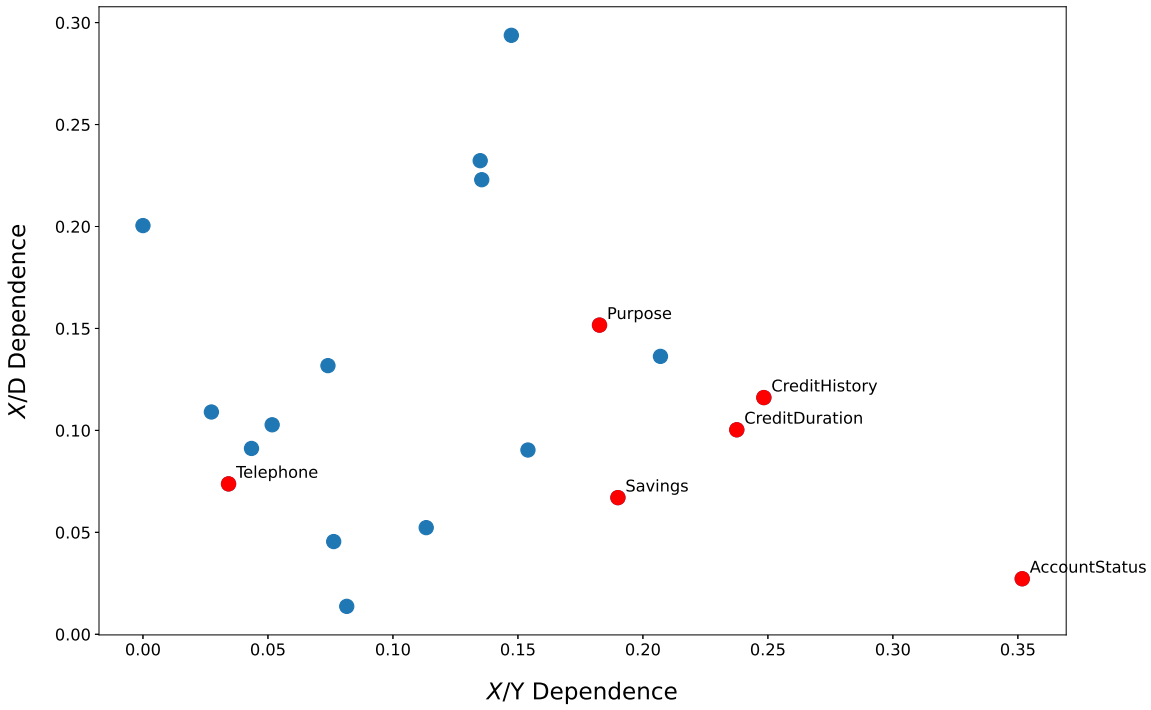
Notes: This figure displays the Cramer's V measures between each feature and the default variable (horizontal axis) and the gender variable (vertical axis).

Figure 4: Fairness PDP for the statistical parity in TREE prime model



Notes: Each subplot displays the FPDP for statistical parity, associated to a given feature and the classification TREE-prime model with indirect discrimination. The Y-axis displays the p-value of the statistical parity test statistic. Red line represents the 10% threshold.

Figure 5: Measures of association between features, target variables, and gender



Notes: This figure displays the dependence of each feature with respect to the target (horizontal axis) and the gender (vertical axis), using Cramer's v as a measure of dependence. Red dots correspond to candidate variables according to the statistical parity test applied to the TREE-prime model.

Table 1: Model performances with and without the protected feature

Panel A: Models with gender							
	LR	LR (Ridge)	TREE	RF	XGB	SVM	ANN
PCC	77.4	76.6	77.3	87.3	81.3	78.2	79.1
AUC	0.8279	0.8268	0.8266	0.938	0.8877	0.8107	0.8341
Panel B: Models without gender							
	LR	LR (Ridge)	TREE	RF	XGB	SVM	ANN
PCC	77.2	76.7	81.5	87.4	79.6	76.0	81.1
AUC	0.8264	0.8200	0.8866	0.9372	0.8261	0.8059	0.8754

Notes: This table reports the percentage of correct classification (PCC) and the area under the ROC curve (AUC) for each scoring model, with gender (Panel A) and without gender (Panel B). LR: logistic regression, LR(Ridge): logistic ridge regression, TREE: classification tree, RF: random forest, XGB: XGBoosting, SVM: support vector machine, ANN: artificial neural network.

Table 2: Fairness tests for models with gender

	LR	Ridge	TREE	RF	XGB	SVM	ANN
Statistical parity	0.0003*	0.0001*	0.0097*	0.0349*	0.0000*	0.0041*	0.0041*
Cond. parity Group 1	0.0003*	0.0000*	0.0035*	0.0214*	0.0000*	0.0008*	0.0036*
Cond. parity Group 2	0.0719	0.0986	0.4909	0.3226	0.0331*	0.3223	0.0395*
Cond. parity (global)	0.0003*	0.0000*	0.0110*	0.0434*	0.0000*	0.0022*	0.0017*
Equal odds	0.0185*	0.0039*	0.2387	0.8220	0.0004*	0.1436	0.0802
Equal opportunity	0.0888	0.0344*	0.3012	0.7796	0.0004*	0.1675	0.6554
Predictive equality	0.0242*	0.0100*	0.1801	0.5753	0.0945	0.1598	0.0277*

Notes: This table reports the p-values of the fairness tests (chi-squared) obtained for the scoring models using gender as an explanatory variable. * indicates statistical significance at 5%. LR: logistic regression, LR(Ridge): logistic ridge regression, TREE: classification tree, RF: random forest, XGB: XGBoosting, SVM: support vector machine, ANN: artificial neural network.

Table 3: Fairness tests for models without gender

	LR	Ridge	TREE	RF	XGB	SVM	ANN
Statistical parity	0.0734	0.1110	0.5310	0.1206	0.0965	0.2913	0.0067*
Cond. parity Group 1	0.0989	0.0304*	0.5950	0.0966	0.0431*	0.1693	0.0072*
Cond. parity Group 2	0.0866	0.6874	0.2130	0.3226	0.3531	0.8506	0.0631
Cond. parity (global)	0.0590	0.0885	0.3998	0.1542	0.0841	0.3821	0.0048*
Equal odds	0.6712	0.4196	0.5645	0.9242	0.7202	0.6754	0.1727
Equal opportunity	0.7746	0.7209	0.8892	0.7796	0.4213	0.5175	0.6602
Predictive equality	0.3977	0.2046	0.2890	0.7783	0.9216	0.5451	0.0685

Notes: This table reports the p-values of the fairness tests (chi-squared) obtained for the scoring models estimated without the gender variable. * indicates statistical significance at 5%. LR: logistic regression, LR(Ridge): logistic ridge regression, TREE: classification tree, RF: random forest, XGB: XGBoosting, SVM: support vector machine, ANN: artificial neural network.

Table 4: Fairness tests for the TREE models

	TREE	TREE-prime	TREE-modif
Statistical parity	0.5310	0.0216*	0.5195
Cond. parity Group 1	0.5950	0.0552	0.7849
Cond. parity Group 2	0.2130	0.0305*	0.0973
Cond. parity (global)	0.3998	0.0153*	0.2438
Equal odds	0.5645	0.0363*	0.3441
Equal opportunity	0.8892	0.0101*	0.4547
Predictive equality	0.2890	0.8852	0.2095
PCC	81.5	79.0	77.8
AUC	0.8866	0.8393	0.8345

Notes: This table reports the p-values of the fairness tests obtained for three different decision trees. The first one (TREE) corresponds to the decision tree without the gender variable. TREE-prime denotes a decision tree obtained with the same feature space, but with an alternative hyperparameter tuning strategy. TREE-modif is a modified version of the previous one for which we retrieve the telephone variable from the decision rules. * indicates statistical significance at 5%.

A Appendix

A.1 Fairness tests

As an example, we detail the notations for the standard chi-squared conditional independence tests. Denote by A, B, C the variables of interest for a given fairness definition, with $A \in \{a_1, a_2\} = \{0, 1\}$ and $B \in \{b_1, b_2\} = \{0, 1\}$ the two binary variables for which we test independence for, and $C \in \{c_1, ..., c_k, ..., c_K\}$ the conditioning variables. For instance, testing the null hypothesis $\mathrm{H}_{0,EO}$ of equal odds implies $A = \widehat{Y}$, $B = D$, and $C = Y$. This notation encompasses most of the fairness metrics used in the literature, such as statistical parity ($A = \widehat{Y}$, $B = D$, and $C = \varnothing$), conditional statistical parity ($A = \widehat{Y}$, $B = D$, and $C = X_c$), predictive equality ($A = \widehat{Y}$, $B = D$, and $C = (Y = 1)$), and equal opportunity ($A = \widehat{Y}$, $B = D$, and $C = (Y = 0)$).

Let us denote by n_{uvk} the number of times outcome $A = a_u$, $B = b_v$, and $C = c_k$ is observed over the n instances, with $\sum_{u,v,k} n_{uvk} = n$. The vector $\boldsymbol{n} = (n_{111}, \ldots, n_{22K})$ follows a multinomial distribution with parameters n and $\boldsymbol{p} = (p_{111}, \ldots, p_{22K})$, the vector of coresponding probabilities $p_{uvk} \geq 0$ with $\sum_{u,v,k} p_{uvk} = 1$. The marginal distributions are defined by using the symbol "+" which refers to summation over a subscript. Hence, p_{u++} , p_{+v+}, and p_{++k} refers respectively to the marginal probabilities $\Pr(A = a_u)$, $\Pr(B = b_u)$, and $\Pr(C = c_k)$. Similarly, $p_{uv+} = \Pr(A = a_u, B = b_v)$ denotes the joint probability of A and B. Within this framework, we can use a standard Pearson chi-squared test for each conditioning event $C = c_k$. The corresponding test statistic is defined as:

$$\chi^2_{(k)} = \sum_u \sum_v \frac{\left(n_{uvk} - \mathbb{E}\left(n_{uvk}\right)\right)^2}{\mathbb{E}\left(n_{uvk}\right)} \tag{11}$$

with $\mathbb{E}\left(n_{uvk}\right) = n_{u+k} \times n_{+vk} / n_{++k}$, where n_{u+k} denotes the number of times outcome $A = a_u$ and $C = c_k$ is observed over the n instances, whatever the observed value of B, n_{+vk} denotes the number of times outcome $B = b_v$ and $C = c_k$ is observed whatever the observed value of A, and n_{++k} is number of outcomes with $C = c_k$. The test statistic $\chi^2_{(k)}$ for each class c_k has an approximate chi-squared distribution with one degree of freedom. Thus, under the null of fairness $H_{0,i}$, the global test statistic $F_{H_{0,i}} = \sum_k \chi^2_{(k)}$ has a chi-squared distribution with K degrees of freedom. For a significance level $\alpha \in]0, 1/2[$, the null hypothesis $\mathrm{H}_{0,i}$ is rejected as soon as $\chi^2 > d_{1-\alpha}$, where $d_{1-\alpha}$ denotes the $1 - \alpha$ quantile of the $\chi^2(K)$ distribution.

A.2 Database description

Table 5: Database description

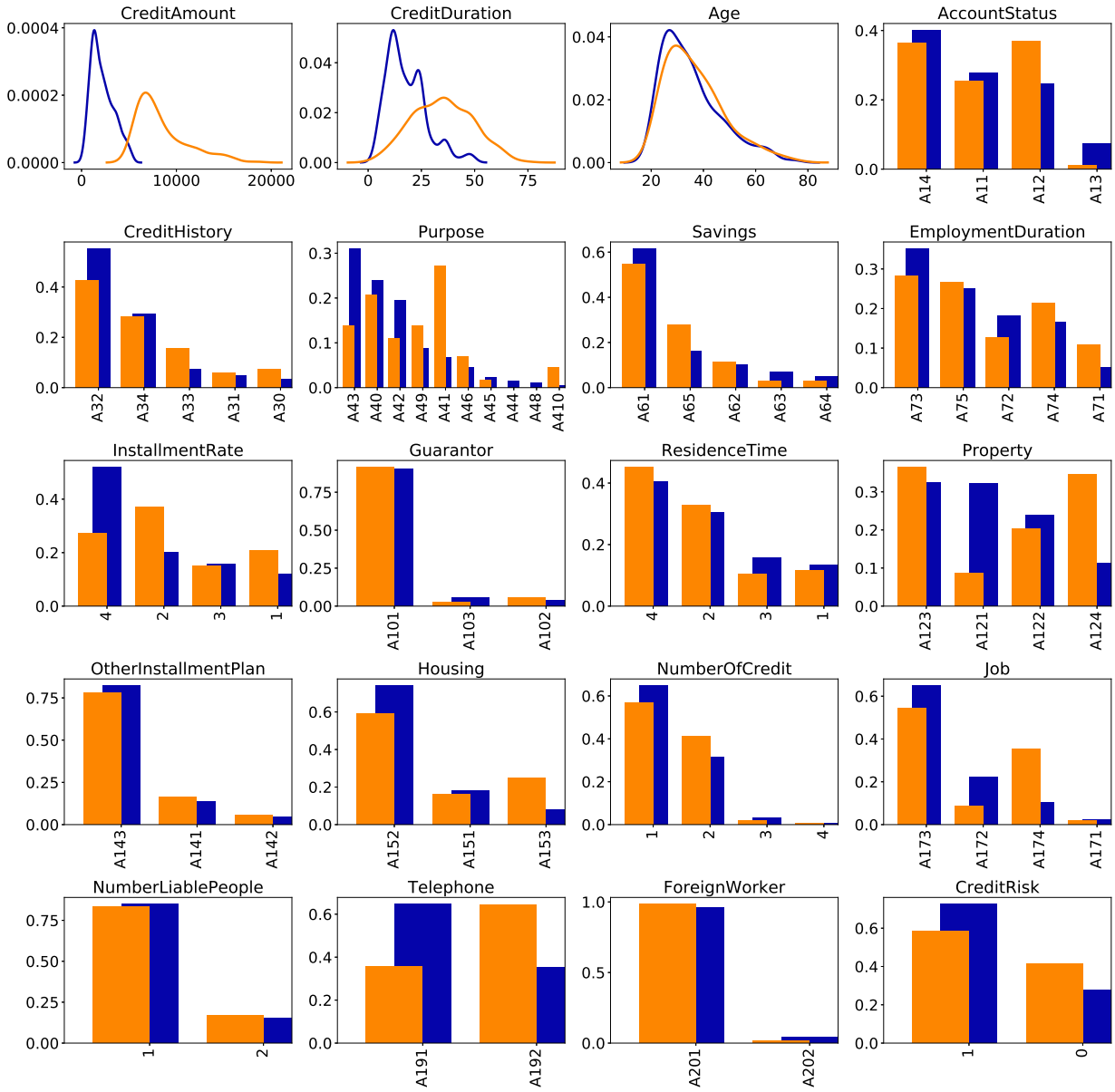
Short name	Complete name	Variable type	Domain
Age	Age	Numerical	$\mathbb{R}^+$
CreditAmount	Credit amount	Numerical	$\mathbb{R}^+$
CreditDuration	Credit duration	Numerical	$\mathbb{R}^+$
AccountStatus	Status of existing checking account	Categorical	#4
CreditHistory	Credit history	Categorical	#5
Purpose	Credit Purpose	Categorical	#10
Savings	Status of savings accounts and bonds	Categorical	#5
EmploymentDuration	Employment length	Categorical	#5
InstallmentRate	Installment rate	Numerical	$\{1, 2, 3, 4\}$
Gender&PersonalStatus	Personal status and gender	Categorical	#4
Guarantor	Other debtors	Categorical	#3
ResidenceTime	Period of present residency	Numerical	$\{1, 2, 3, 4\}$
Property	Property	Categorical	#4
OtherInstallmentPlan	Installment plans	Categorical	#3
Housing	Residence	Categorical	#3
NumberOfCredit	Number of existing credits	Numerical	$\{1, 2, 3, 4\}$
Job	Employment	Categorical	#4
NumberLiablePeople	Dependents	Numerical	$\{1, 2\}$
Telephone	Telephone	Binary	#2
ForeignWorker	Foreign worker	Binary	#2
CreditRisk	Credit score	Binary	#2

Table 6: Feature overview

Complete name	Description
Age	Age in years
Credit amount	Credit amount
Credit duration	Duration in month
Status of existing checking account	A11 : ... $<$ 0 DM, A12 : 0 $<=$ ... $<$ 200 DM A13 : $\geq$ 200 DM / salary assignments (1 year) A14 : no checking account
Credit history	A30: no credits taken/ all credits paid back duly A31: all credits at this bank paid back duly A32: existing credits paid back duly till now A33: delay in paying off in the past A34: other credits existing (not at this bank)
Credit Purpose	A40: car (new), A41: car (used), A42: equipment A43: radio/television, A44: domestic appliances A45: repairs, A46: education, A48: retraining A49: business, A410: others
Status of savings accounts and bonds	A61: $<$ 100 DM, A62: 100 $\leqslant x <$ 500 A63: 500 $\leqslant x <$ 1000, A64: $\geq$ 1000 DM A65: unknown/ no savings account
Employment duration	A71: unemployed, A72: . $<$ 1 year A73: 1 $\leqslant x <$ 4 years, A74 : 4 $\leqslant x <$ 7 years, A75: $\geq$ 7 years
Installment rate	Installment rate in percentage of disposable income
Personal status and gender	A91: male : divorced/separated A92: female : divorced/separated/married A93: male : single, A94: male : married/widowed
Other debtors	A101: none, A102: co-applicant
Period of present residency	Present residence since
Property	A121: real estate, A123: car or other, A122: building society savings agreement A124: unknown / no property
Installment plans	A141: bank, A142: stores, A143: none
Housing	A151: rent, A152: own, A153: for free
Number of existing credits	Number of existing credits at this bank
Employment	A171: unemployed/ unskilled - non-resident A172: unskilled - resident A173: skilled employee A174: management/ self-employed/highly qualified
Dependents	Number of people being liable to provide maintenance
Telephone	A191: none, A192: yes
Foreign worker	A201: yes, A202: no
Credit score	1: Good, 2: Bad

A.3 Feature distribution by class of risk

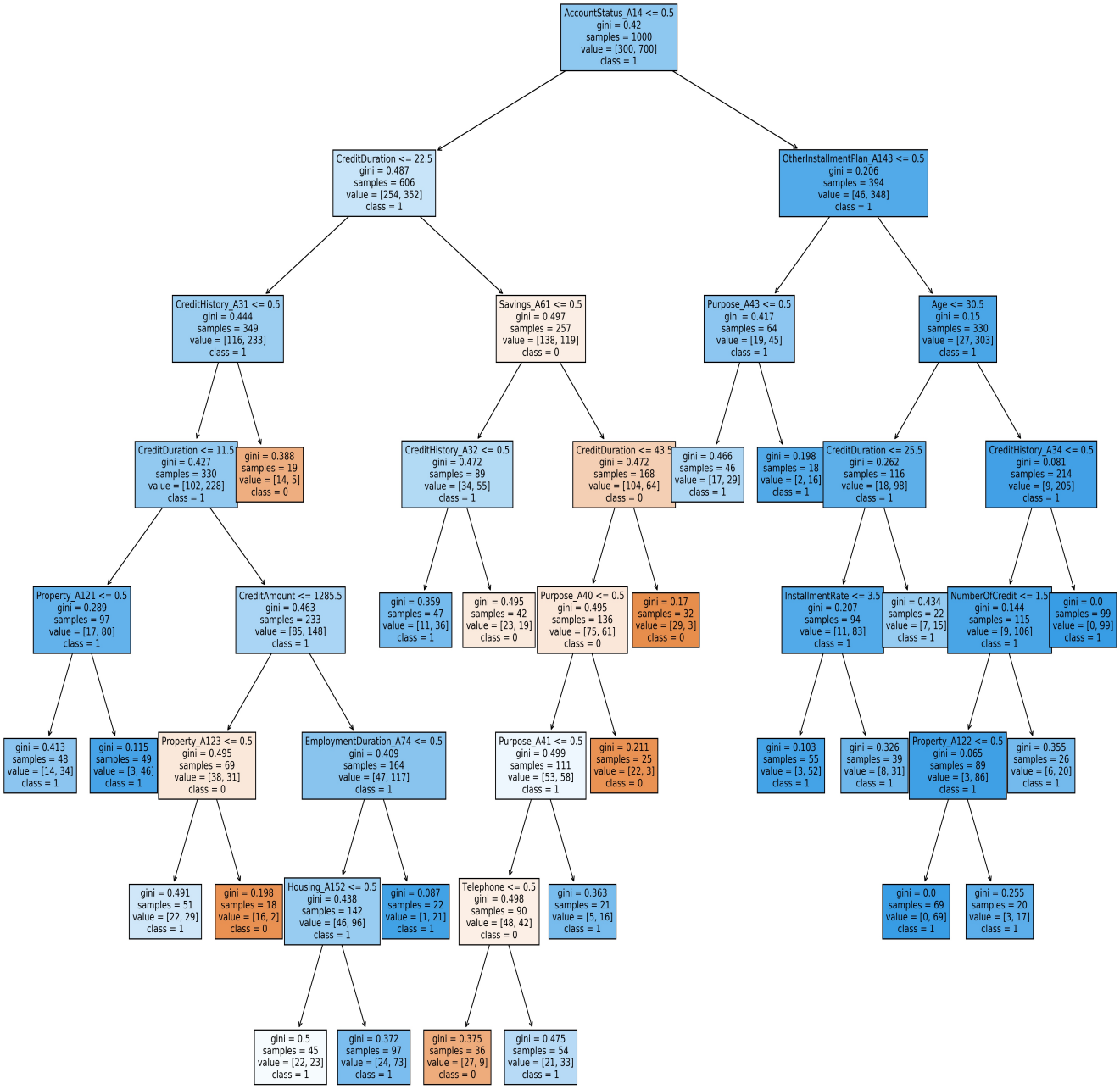
Figure 6: Feature distribution by class of risk



Notes: These figures display the feature distributions by class of risk, using kernel density estimation for continuous variables. Blue color refers to Group 1 (low-risk profiles) and orange to Group 2 (high-risk profiles).

A.4 Decision tree

Figure 7: Decision tree for the TREE-prime model



Notes: This figure displays the decision tree obtained with the hyperparameters mentionned in column "TREE-prime" of table 7 and by excluding the gender variable from the features.

A.5 Hyperparameters of machine-learning models

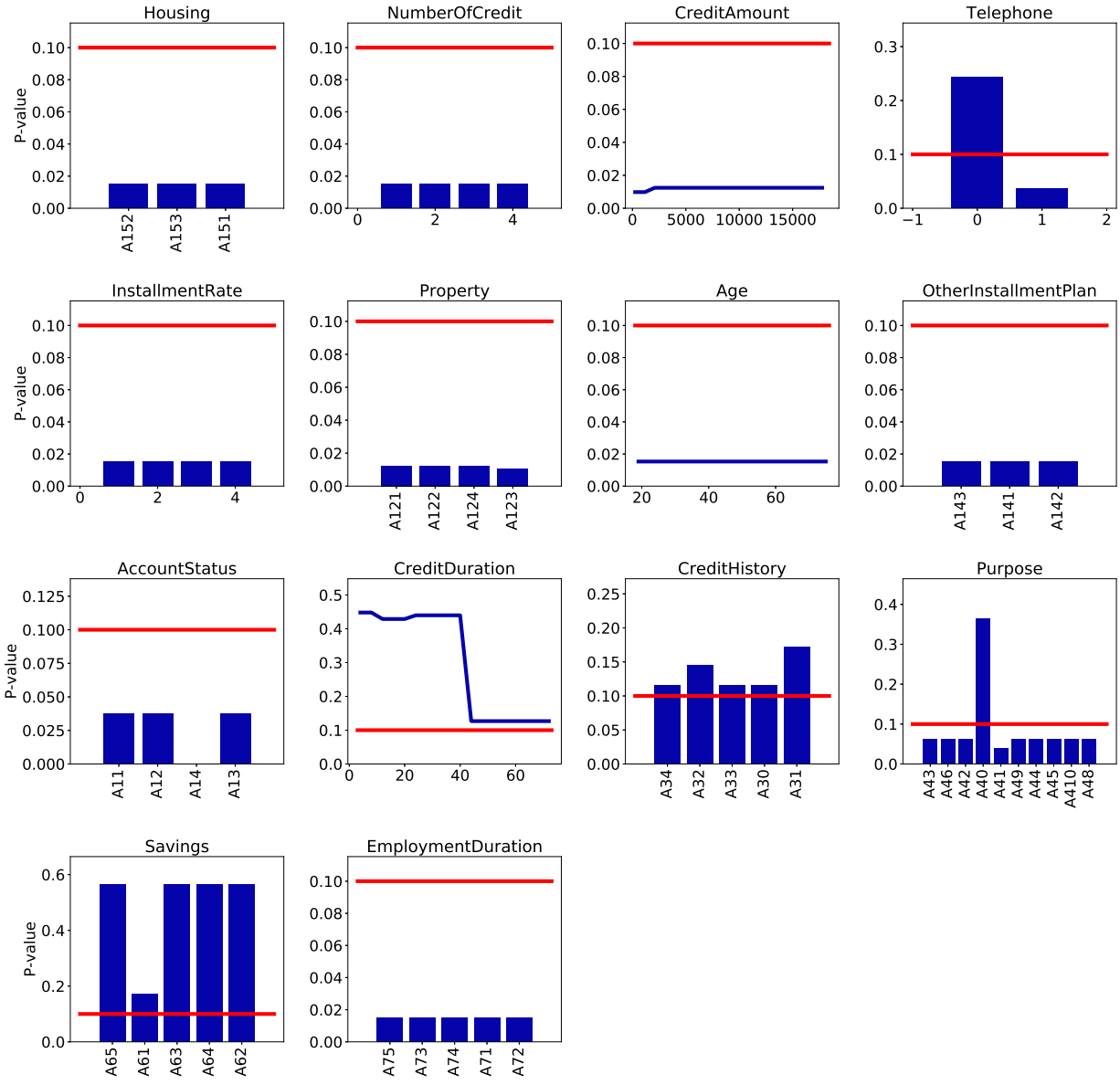
Table 7: Hyperparameter tuning

Hyperparameter set	RIDGE	TREE	TREE prime	ANN	SVM	RF	XGBoost
Criterion		Entropy, Gini (Gini)	Entropy, Gini (Gini)			Entropy, Gini (Entropy)	
Max. depth of the tree		1-29 (20)	1-9 (7)			1-99 (82)	1-5 (2)
Min. number of individuals required to split a node		2-9 (2)	2-59 (56)			2-49 (3)	
Min. number of individuals by leaf		1-19 (5)	1-59 (18)			1-29 (10)	
Number of inputs compared at each split		all, $\sqrt{k}$, $\log 2(k)$ ($\sqrt{k}$)	all, $\sqrt{k}$, $\log 2(k)$ (all)			all, $\sqrt{k}$, $\log 2(k)$ (all)	
Decrease of the impurity greater than or equal to this value		0-0.9 (0)	0-0.9 (0)			0-0.9 (0)	
Optimization method: Grid Search				Yes			
Number of hidden layers				1			
Number of neurons by hidden layer				1-25 (20)			
Max. number of iterations				250			
Activation function				relu			
Solver for weight optimization				adam			
L2 penalty (regularization term) parameter	0.0001-20.0001 (0.2001)			0.0001			
Early stopping				True			
Max. number of epochs to not meet tol improvement				50			
Regularization parameter					1		
Kernel					linear		
Learning rate							0-0.9 (0.5)
Min. sum of instance weight (hessian) needed in a child							1-39 (15)
Min. loss reduction required to make a further partition of a leaf node							0-0.9 (0.3)
Subsample ratio of columns when constructing each tree							0-0.9 (0.4)
Subsample ratio of columns for each level							0-0.9 (0.1)
Subsample ratio of columns for each node (split)							0-0.9 (0.2)

Notes: This table displays the hyperparameter tuning procedures for the various machine-learning models used in the numerical analysis: Decision Tree (Tree), Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), and XGBoost. The values in parentheses are the optimal hyperparameter values.

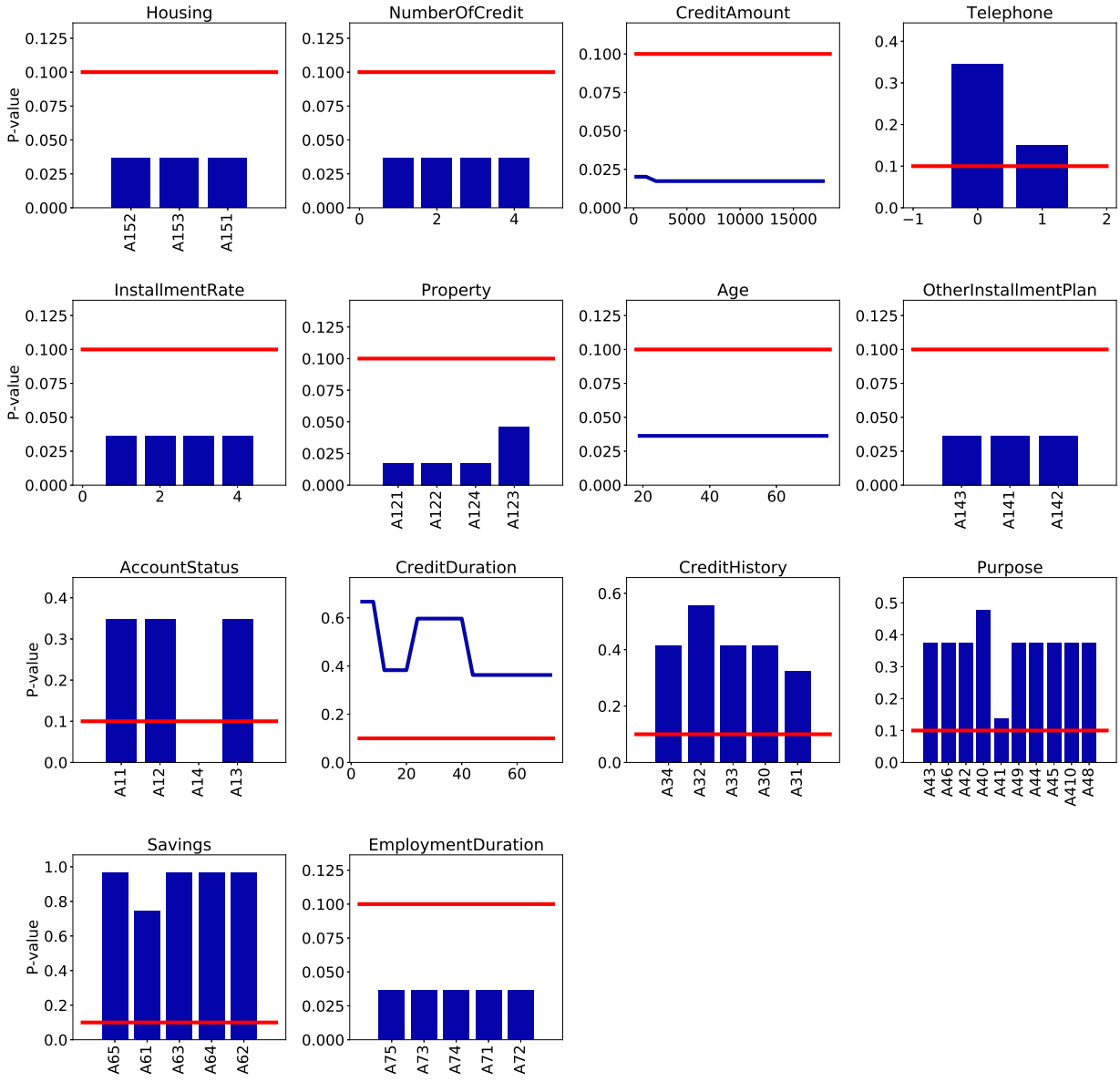
A.6 FPDP analysis for TREE-prime model

Figure 8: Fairness PDP for conditional statistical parity in TREE prime model



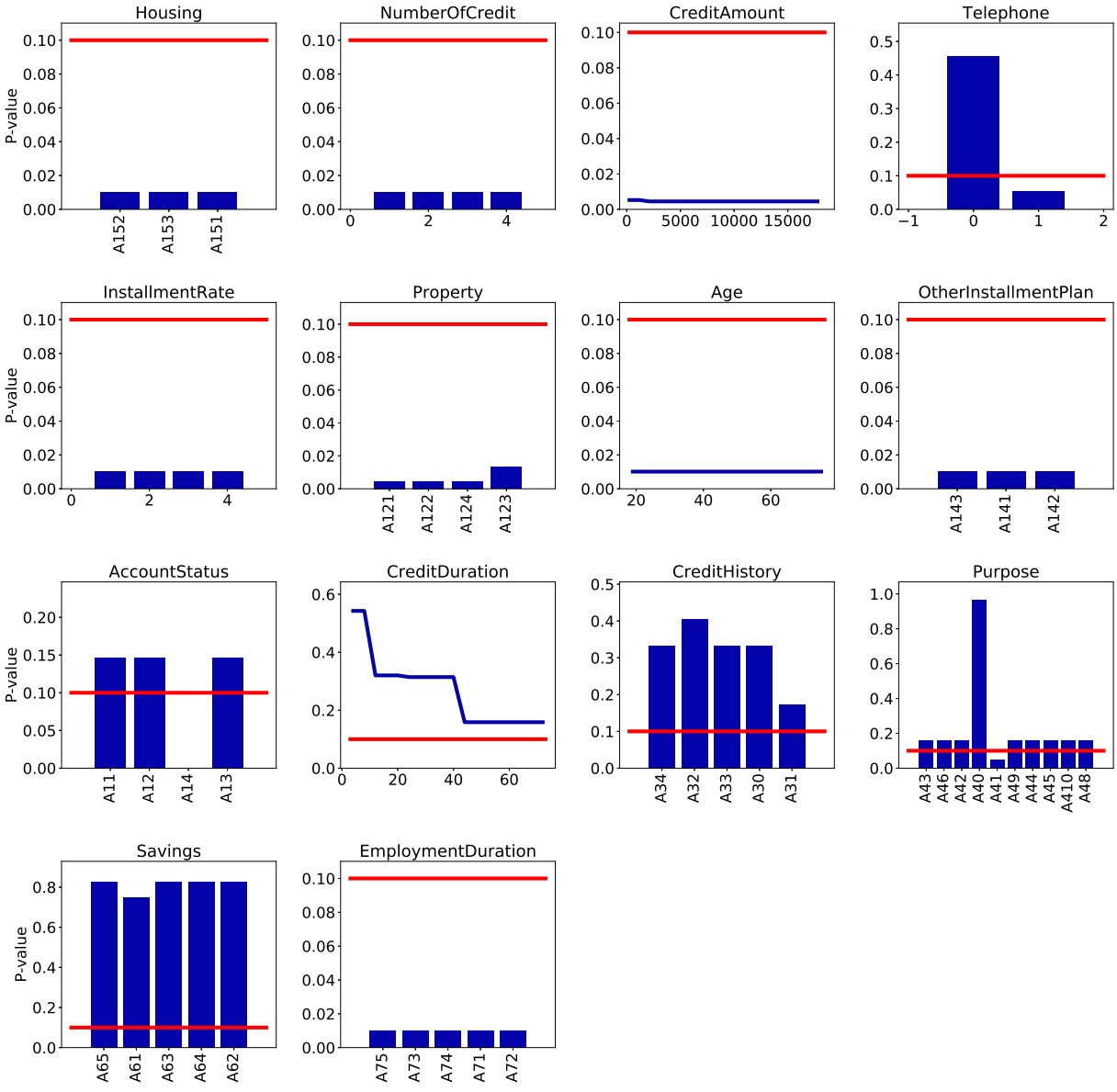
Notes: Each subplot displays the FPDP for conditional statistical parity, associated to a given feature and the classification TREE-prime model with indirect discrimination. The Y-axis displays the p-value of the conditional statistical parity statistic. Red line represents the 10% threshold.

Figure 9: Fairness PDP for equal odds in TREE prime model



Notes: Each subplot displays the FPDP for equal odds, associated to a given feature and the classification TREE-prime model with indirect discrimination. The Y-axis displays the p-value of the equal odds test statistic. Red line represents the 10% threshold.

Figure 10: Fairness PDP for equal opportunity in TREE prime model



Notes: Each subplot displays the FPDP for equal opportunity, associated to a given feature and the classification TREE-prime model with indirect discrimination. The Y-axis displays the p-value of the equal opportunity. Red line represents the 10% threshold.