The Fairness of Credit Scoring Models

CRC Seminar

Christophe Hurlin (LEO), Christophe Pérignon (HEC) and Sébastien Saurin (LEO)

University of Orléans

July 9, 2021

Introduction

Why is AI so popular in credit markets?

- Allow processing massive quantity of data using powerful algorithms.
- Associated reduction in Type-1 and Type-2 errors benefits both banks (profit) and borrowers (access, rate).
- Speed up and lower the cost of loan application processing, hence including small borrowers traditionally overlooked by standard screening techniques (Berg et al. (RFS 2020)).
- Lead to higher responsiveness of the credit supply to demand shocks and lower interest rates (Fuster et al. (RFS, 2019)).
- Favour the emergence of new, less regulated players (Fintech), granting loans to high risk-weight borrowers.

Are algorithms neutral?

AI has also stirred a passionate debate about the associated potential **discrimination biases** (O'Neil (2017), Bartlett et al. (2021)).

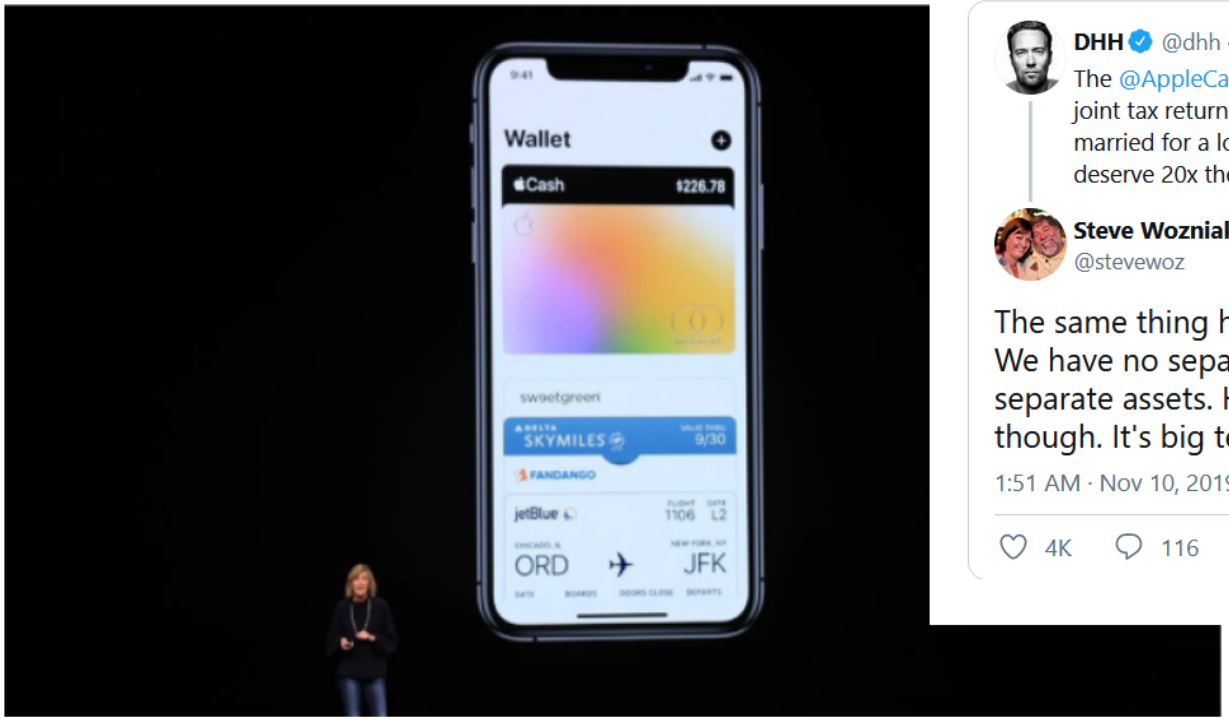
- Credit scoring algorithms can place groups of individuals sharing a protected attribute, such as **gender**, age, citizenship or race, at a systematic disadvantage (rejection rate or interest rate).
- Using data on US mortgages, **Fuster et al.** (RFS 2021) found that the use of ML algorithm increases interest rate disparity between White/Asian borrowers and Black/Hispanic borrowers.
- Algorithmic discrimination 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**.

Introduction

The New York Times

Apple Card Investigated After Gender Discrimination Complaints

A prominent software developer said on Twitter that the credit card was "sexist" against women applying for credit.



DHH @dhh · Nov 7, 2019

The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.

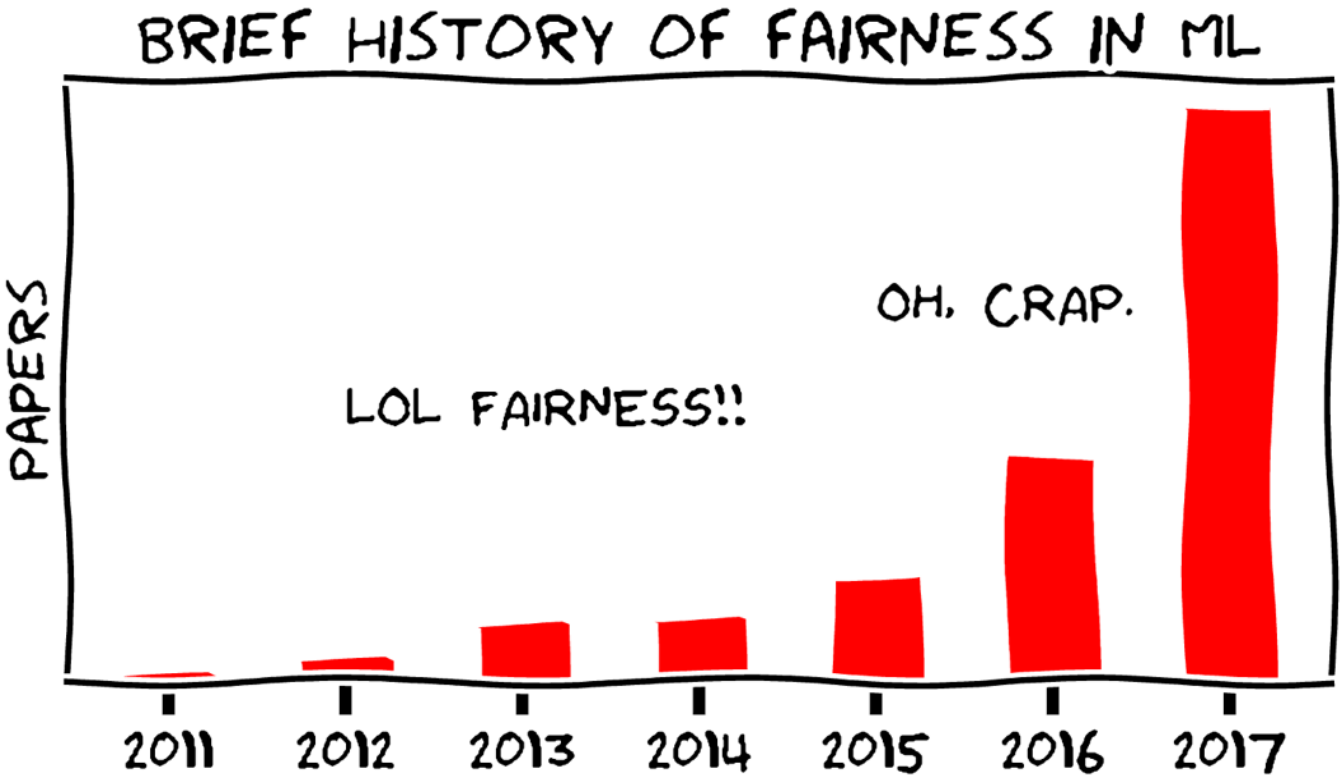
Steve Wozniak
@stevewoz

The same thing happened to us. I got 10x the credit limit. We have no separate bank or credit card accounts or any separate assets. Hard to get to a human for a correction though. It's big tech in 2019.

1:51 AM · Nov 10, 2019

4K 116 Copy link to Tweet

Introduction



Introduction

EUROPEAN
COMMISSION

Brussels, 21.4.2021
COM(2021) 206 final

2021/0106 (COD)

Proposal for a

REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS

{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

Introduction

A concern for the regulator

The recent proposal for 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*"

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

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.

Sources of discrimination

1. 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.

2. The scoring algorithm can learn from a dataset of actual defaults that occurred in the past. Two cases arise:

    1. **Disparate treatment (direct discrimination)** occurs when the algorithm explicitly uses the protected attribute.

    2. **Disparate impact (indirect discrimination, proxy discrimination)** occurs when the lender use facially neutral variables that are able, collectively, to synthetically reconstruct the protected attribute.

In practice, disparate impact is more likely to occur with highly non-linear, non-interpretable, and opaque scoring models, such as advanced ML algorithms.

Goal of the paper

We design and implement a methodology allowing algorithmic lenders, as well as their regulators, to address this question in three steps.

1. First, we quantify the difference of treatment across groups and **formally test** whether we can reject the null hypothesis of fairness.
2. We propose a novel and simple **interpretability technique**, called *Fairness Partial Dependence Plot* (FPDP), to identify the variables that cause the lack of fairness.
3. Third, once the **candidate variables** have been identified, we check whether all these variables are legitimate in the lending context.

Application

We illustrate our fairness assessing framework by testing for gender discrimination in a database of retail borrowers.

- We show that our tests are able to detect direct discrimination when gender is explicitly used as a feature to assess creditworthiness.
- 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.
- We show that the choice of the parameters controlling the learning process of the algorithms strongly impacts fairness.
- We show that the FPDP identify the set of candidate features. Reassuringly, the selected features make sense as they play a key role in the algorithm's decision rules.

Introduction

The outline:

Section 2: Fairness tests

Section 3: Fairness interpretability

Section 4: Empirical application

Fairness test

Notations

We consider a bank using an algorithm to screen borrowing applications.

- $Y \in \mathcal{Y} = \{0, 1\}$: target variable, where the value 1 corresponds by convention to the favorable output
- We will thereafter use the *good type* $(Y = 1)$ vs. *bad type* $(Y = 0)$ terminology
- $X_c \in \mathcal{X}$: non-protected attributes, 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)
- $D \in \mathcal{D} = \{0, 1\}$: protected attribute (e.g. gender), with $D = 1$ the protected class

Scoring model

- The goal of the bank is to build a **scoring model** (classifier) $f(X)$ which maps the observable non-protected features X into a predicted outcome:

$$\hat{Y} = f(X_c) \in \mathcal{Y} = \{0, 1\}$$

- No constraint on the f function: it can be parametric (logistic regression) or not, linear or not (TREE), an individual or ensemble classifier (RF) which combines a set of homogeneous or heterogeneous models, etc.

- $p(X) = \Pr(Y = 1 | X)$ denotes the conditional probability for an applicant of being good type

Fairness test

Fairness definitions

Many fairness criteria have been proposed in computer science, machine learning, criminology, and economics, each aiming to capture different dimensions of fairness

See Hardt et al. (2016), Verma and Rubin (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 **four fairness metrics**:

1. **Statistical parity**
2. **Equal odds**
3. **Predictive parity**
4. **Overall accuracy**

Fairness test

Fairness definitions

Many fairness criteria have been proposed in computer science, machine learning, criminology, and economics, each aiming to capture different dimensions of fairness

See Hardt et al. (2016), Verma and Rubin (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 **four fairness metrics**:

1. **Statistical parity**
2. **Equal odds**
3. **Predictive parity**
4. **Overall accuracy**

Fairness test

Statistical parity

Definition

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

Intuition: all applicants should have an equivalent opportunity to obtain a good outcome from the credit scoring model, regardless of their group membership.

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

Conditional statistical parity

In case of disparate impact, the differences in opportunity may be due to a composition effect, rather than to discrimination. It can be more informative to compare similar applicants with comparable individual features (income, job, diploma, etc.).

Definition

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

Fairness test

Equal odds

Intuition: 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.

Definition

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 $\widehat{Y} \perp\!\!\!\perp D|Y$.

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).

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

Equal odds: relaxation

Equal opportunity implies $\widehat{Y} \perp\!\!\!\perp D \Big| Y = 1$ and the equality of the TPRs for the protected and unprotected groups:

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

Predictive equality, reflects the equality of the FPRs for both groups and corresponds to the assumption $\widehat{Y} \perp\!\!\!\perp D | Y = 0$.

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

Fairness test

Fact (A trick for fairness diagnosis)

As all fairness definitions are based on the joint probability of $(\widehat{Y}, Y, D)$ *and can be expressed in terms of independence assumptions, we can derive* **simple inference model-agnostic procedure**.

Statistical parity	$\widehat{Y} \perp\!\!\!\perp D$
Conditional statistical parity	$\widehat{Y} \perp\!\!\!\perp D \vert X_c$
Equal odds	$\widehat{Y} \perp\!\!\!\perp D \vert Y$
Equal opportunity	$\widehat{Y} \perp\!\!\!\perp D \vert Y = 1$
Predictive equality	$\widehat{Y} \perp\!\!\!\perp D \vert Y = 0$

Remark: We do not need to have access to the algorithm but just to its outputs and the protected attribute

Fairness test

Consider a 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.

Fairness test

Definition

A **fairness test statistic** is defined as:

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

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.

Under the null hypothesis $\mathrm{H}_{0,i}$, we 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.

Fairness test

The notation for F_i encompasses a wide class of test statistics:

- Chi-squared conditional independence test
- Cochran–Mantel–Haenszel (CMH) test
- z-tests associated to the hypothesis tests of the difference, ratio, or odds ratio of two independent proportions
- Likelihood ratio tests.
- etc.

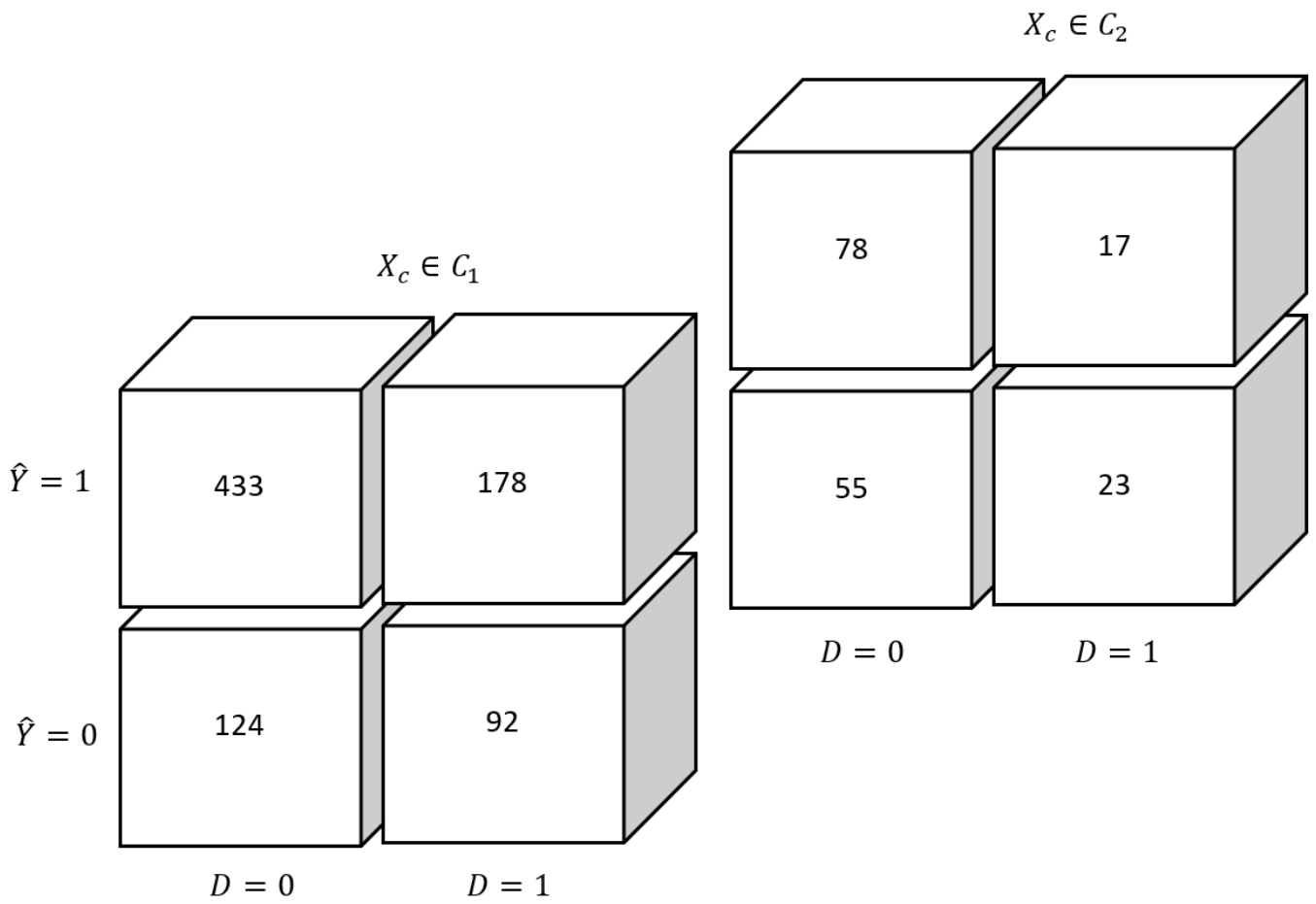
Fairness test

Numerical illustration: chi-squared test of conditional independence

- A regulator whishes to test whether a scoring model satisfies the null hypothesis of conditional statistical parity $H_{0,CSP}$
- Consider a sample of $1,000$ applicants, 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.
- Use a standard Pearson chi-squared (conditional) independence test-statistic

Fairness test

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



Fairness test

Numerical illustration: chi-squared test of conditional independence (cont'd)

- Class 1 ($n_1 = 827$): $\chi^2_1 = 13.15$, $d^2_{1,0.95} = 3.84$,=> reject $\widehat{Y} \perp\!\!\!\perp D$ at 5% for C_1
- Class 2 ($n_2 = 173$): $\chi^2_1 = 3.24$, $d^2_{1,0.95} = 3.84$ => do not reject $\widehat{Y} \perp\!\!\!\perp D$ at 5% for C_2
- Conditional test statistic $F_{H_{0,CSP}} = \chi^2_1 + \chi^2_2 \sim \chi^2_{(2)}$ and $16.39 > d^2_{2,0.05} = 5.99$ => we reject the null hypothesis $H_{0,CSP}$ of conditional statistical parity, i.e., $\widehat{Y} \perp\!\!\!\perp D | X_c$.

Fact

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.*

Fairness interpretability

Let us assume that the null of fairness is rejected. What is the source of discrimination?

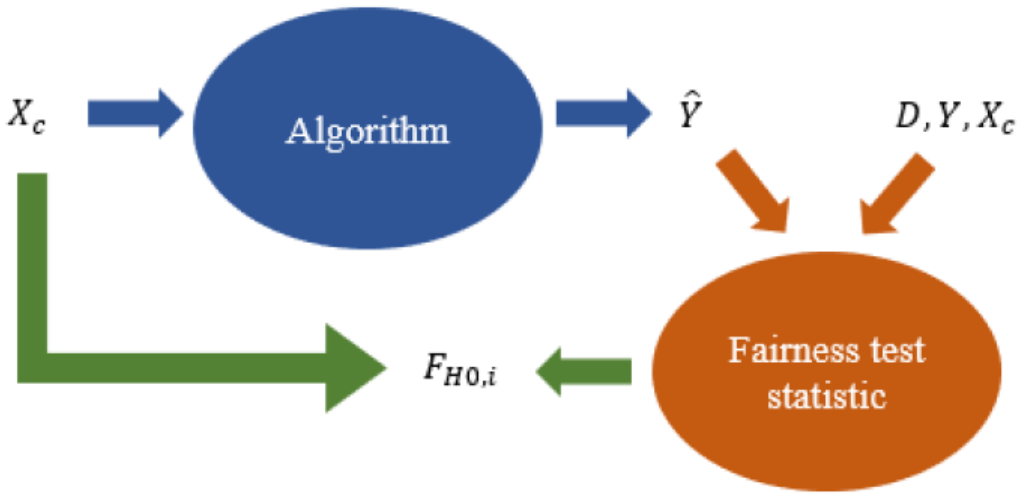
Fairness interpretability

Fairness interpretability

Definition

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 protected attribute induced by a scoring model.

Figure: Fairness interpretability



Interpretability vs. Fairness interpretability

- In general in ML, **interpretability** refers to the capacity of a human to understand the cause of a decision $\widehat{Y}$.
- Interpretability is essential for the governance of AI, especially in the credit scoring industry.
- But, **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.

Fairness interpretability

Interpretation methods

There are different ways to interpret machine learning models and their decisions.

- **Intrinsic interpretability** refers to ML models that are considered interpretable due to their simple structure, such as short 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.
- 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 predicted outcome $\widehat{Y}$ of a model.

Fairness Partial Dependence Plot (FPDP)

Here, we propose a simple model-agnostic method which we call **Fairness Partial Dependence Plot (FPDP)**.

- FPDP displays the marginal effect of a 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: these variables are called **candidate features**.

Fairness interpretability

Fairness Partial Dependence Plot (FPDP)

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

$$F_{H_{0,i}} \equiv h_i(\widehat{Y}_j, Y_j, D_j; j = 1, \ldots, n) = \widetilde{h}_i(X_{A,j}, X_{B,j}, Y_j, D_j; j = 1, \ldots, n)$$

Fairness interpretability

Definition (Fairness Partial Dependence Plot)

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

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

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

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

Fairness interpretability

Consider a case for which the null of fairness is initialy rejected as $F_{H_{0,i}} > d_{1-\alpha}$. A feature is identified as a candidate, if changing its value reverses the fairness diagnosis.

Definition

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 $F_{H_{0,i}}(x) < d_{1-\alpha}$.

Fairness Partial Dependence Plot (FPDP)

The FPDP has two main advantages.

- 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.
- FPDP breaks correlations among features. Indeed, in case of indirect discrimination, some features are correlated with the protected attribute and leads to significantly different outcomes for protected and unprotected groups.

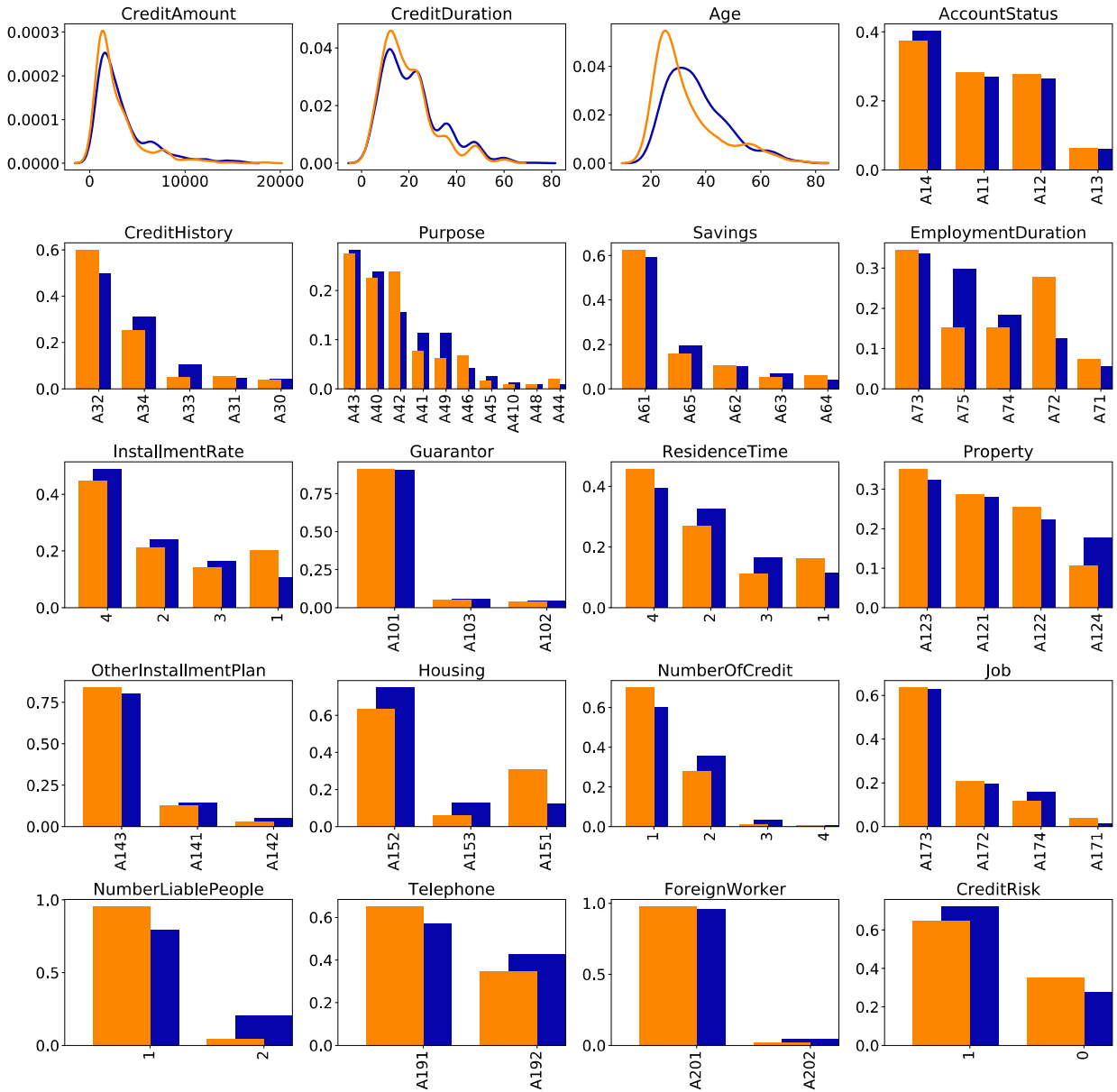
Trick: **breaking the correlations between the features** and **the protected attribute**, and/or **among these proxy features** allows us to identify the candidate variables, as it induces a change in the fairness diagnosis.

Application

Empirical application

German Credit Dataset, Dua and Graff (2019)

- 1,000 consumer loans extended to respectively 310 women and 690 men.
- 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 ($Y = 0$).
- 300 borrowers are in default ($Y = 0$) among which 191 are men and 109 women.
- 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).


CreditAmount
CreditDuration
Age
AccountStatus
CreditHistory
Purpose
Savings
EmploymentDuration
InstallmentRate
Guarantor
ResidenceTime
Property
OtherInstallmentPlan
Housing
NumberOfCredit
Job
NumberLiablePeople
Telephone
ForeignWorker
CreditRisk

Application

Scoring models

1. Logistic regression (LR)
2. Penalized logistic regression (Ridge)
3. Classification tree (TREE)
4. Random forests (RF)
5. XGBoost (XGB)
6. Support vector machine (SVM)
7. Artificial neural networks (ANN).

Hyperparameters: ten-fold cross-validation and a random search algorithm.

Application

Empirical strategy

- 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.
- Then, to know whether the scoring models also leads to indirect discrimination, we remove *gender* from the scoring models. We refer these models as **without-models**.

Application

Table: Model performances with and without the protected feature

Panel A: Models with gender							
	LR	LR (Ridge)	TREE	RF	XGB	SVM	ANN
PCC	77.4	76.4	77.3	87.3	82.1	78.2	79.5
AUC	0.8279	0.8191	0.8266	0.938	0.8723	0.8107	0.8411
Panel B: Models without gender							
	LR	LR (Ridge)	TREE	RF	XGB	SVM	ANN
PCC	77.2	75.7	81.5	87.4	85.8	76	77.7
AUC	0.8264	0.8134	0.8866	0.9372	0.9210	0.8059	0.8276

Note: 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 (CART), RF: random forest, XGB: XGBoosting, SVM: support vector machine, ANN: artificial neural network

Application

Fairness tests for with-models

	LR	Ridge	TREE	RF	XGB	ANN
Statistical parity	**0.0003***	**0.0001***	**0.0097***	**0.0349***	**0.0004***	**0.0001***
Cond. parity Group 1	**0.0003***	**0.0001***	**0.0035***	**0.0214***	**0.0004***	**0.0002***
Cond. parity Group 2	0.0719	0.0781	0.4909	0.3226	**0.0417***	**0.0154***
Cond. parity (global)	**0.0003***	**0.0001***	**0.0110***	**0.0434***	**0.0003***	**0.0001***
Equal odds	**0.0185***	**0.0009***	0.2387	0.8220	**0.0208***	**0.0084***
Equal opportunity	0.0888	**0.0105***	0.3012	0.7796	**0.0206***	**0.0342***
Predictive equality	**0.0242***	**0.0060***	0.1801	0.5733	0.1220	**0.0242***

Note: 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 (CART), RF: random forest, XGB: XGBoosting, SVM: support vector machine, ANN: artificial neural network.

Application

Fairness tests for without-models

	LR	Ridge	TREE	RF	XGB	ANN
Statistical parity	0.0734	0.1373	0.5310	0.1206	0.1288	**0.0159***
Cond. parity Group 1	0.0989	0.0513	0.5950	0.0966	0.2042	**0.0086***
Cond. parity Group 2	0.0866	0.4667	0.2130	0.3226	0.0514	0.1445
Cond. parity (global)	0.0590	0.1188	0.3998	0.1542	0.0670	**0.0109***
Equal odds	0.6712	0.8003	0.5645	0.9242	0.8995	0.3245
Equal opportunity	0.7746	0.9042	0.8892	0.7796	0.7157	0.3917
Predictive equality	0.3977	0.5115	0.2890	0.7783	0.7783	0.2180

Note: 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 (CART), RF: random forest, XGB: XGBoosting, SVM: support vector machine, ANN: artificial neural network.

Application

Operational risk

- We illustrate that the choice of the **hyperparameters** used to control the learning process of credit scoring models 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 hyperparameter tuning strategy.

Hyperparameter set	TREE	TREE prime
Criterion	Entropy, Gini (Gini)	Entropy, Gini (Gini)
Maximum depth of the tree	1-29 (20)	1-9 (7)
Minimum number of individuals required to split a node	2-9 (2)	2-59 (56)
Minimum number of individuals by leaf	1-19 (5)	1-59 (18)
Number of inputs compared at each split	all, $\sqrt{k}$, $\log 2(k)$ ($\sqrt{k}$)	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)

Application

Fairness tests for the TREE models

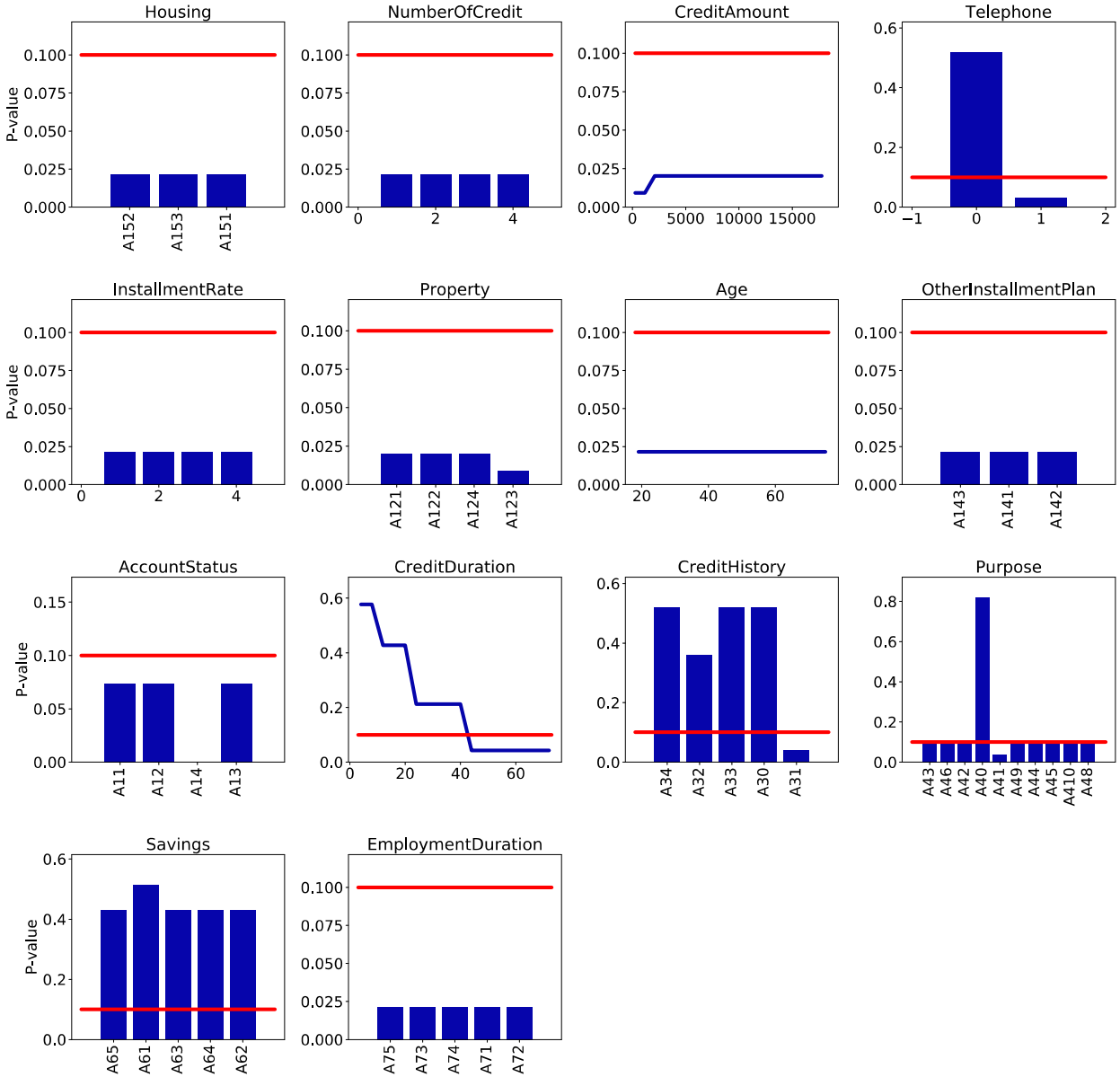
	TREE	TREE-prime	TREE-corr
Statistical parity	0.5320	**0.0216***	0.5195
Cond. parity Group 1	0.5950	0.0552	0.7849
Cond. parity Group 2	0.2130	**0.0305***	0.0973
Cond. parity (global)	0.3998	**0.0153***	0.2438
Equal odds	0.5645	**0.0363***	0.3441
Equal opportunity	0.8892	**0.0101***	0.4547
Predictive equality	0.2890	0.8852	0.2095
PCC	81.5	79.0	77.8
AUC	88.6	83.9	83.4

Note: 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-corr is a modified version of the previous one for which we retrieve the telephone variable from the decision rules. * indicates statistical significance at 5%.

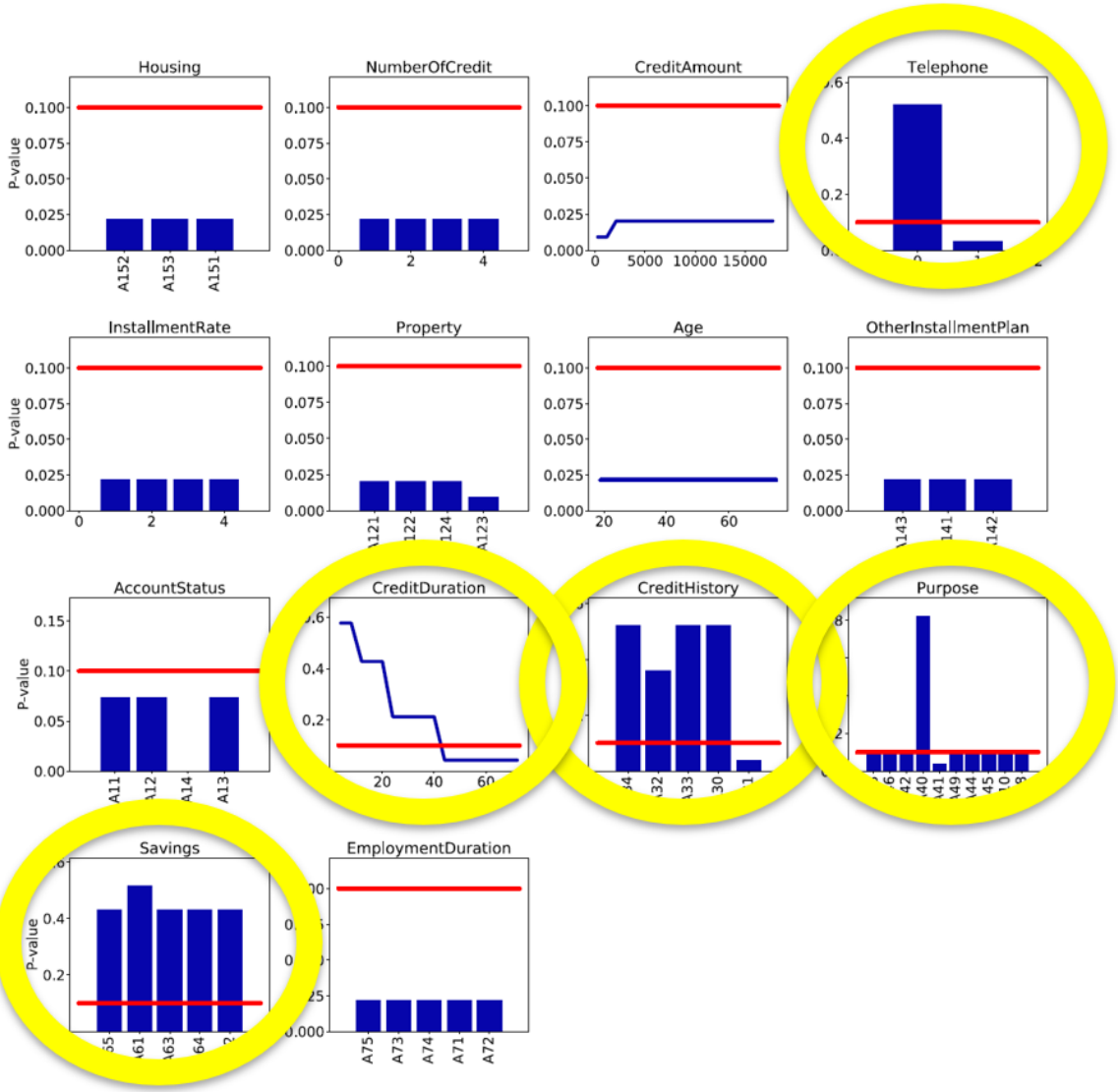
Application

Fairness interpretability

- Several scoring models were classified as unfair by our testing procedure.
- We are now going to identify the **candidate 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.

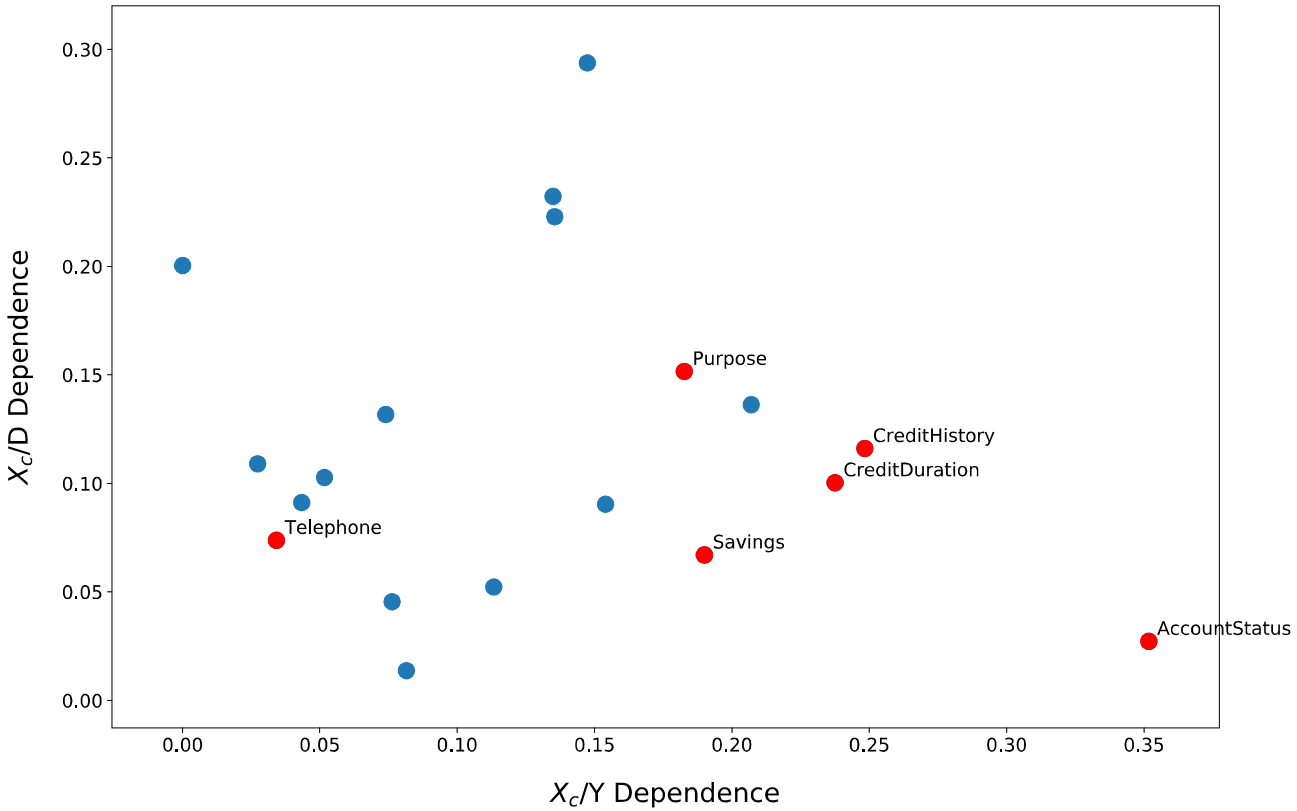

Housing
NumberOfCredit
CreditAmount
Telephone
InstallmentRate
Property
Age
OtherInstallmentPlan
AccountStatus
CreditDuration
CreditHistory
Purpose
Savings
EmploymentDuration
P-value

Application



Application

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



Note: 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

Application

Fairness tests for the TREE models

	TREE	TREE-prime	TREE-corr
Statistical parity	0.5320	**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	88.6	83.9	83.4

Note: 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-corr is a modified version of the previous one for which we retrieve the telephone variable from the decision rules. * indicates statistical significance at 5%.

Conclusion

- 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.