# The Fairness of Credit Scoring Models

C. Hurlin [1] C. Pérignon [2] S. Saurin [1]

[1]University of Orléans

[2]HEC Paris

September 23, 2021

# Why is AI so popular in credit markets?

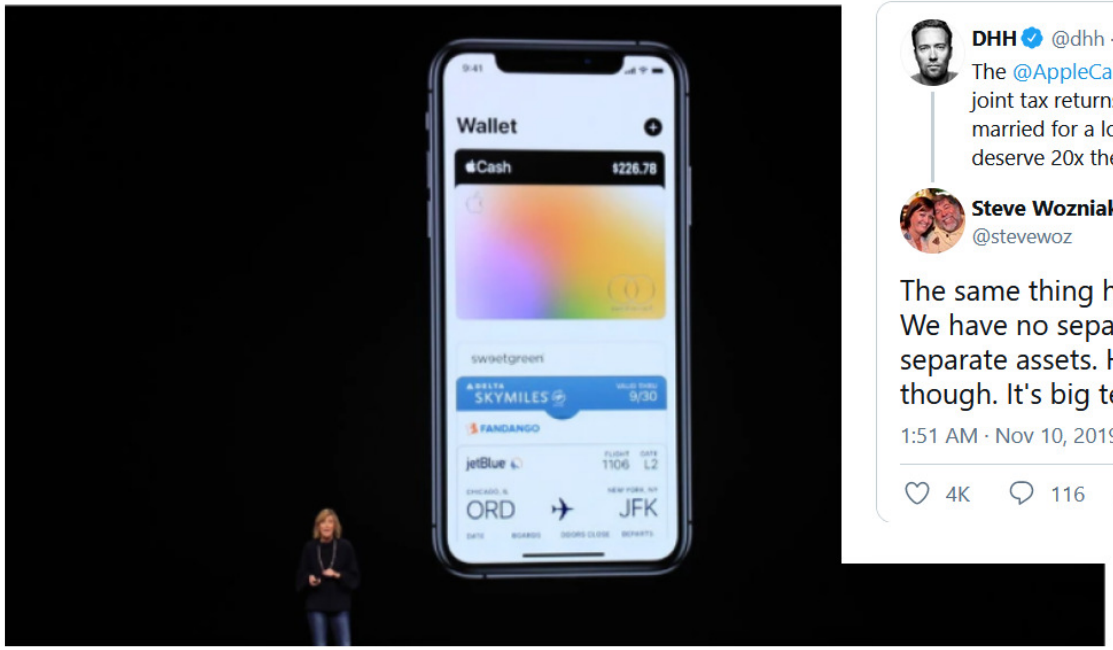
- Massive quantity of data
- Powerful algorithms
- Speed up and lower the cost of loan application processing
- Favour the emergence of **new, less regulated players** (Fintech)
- Unlike humans, algorithms are supposed to be **neutral**.

# Are algorithms really neutral?

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

# E.U. Regulation

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**

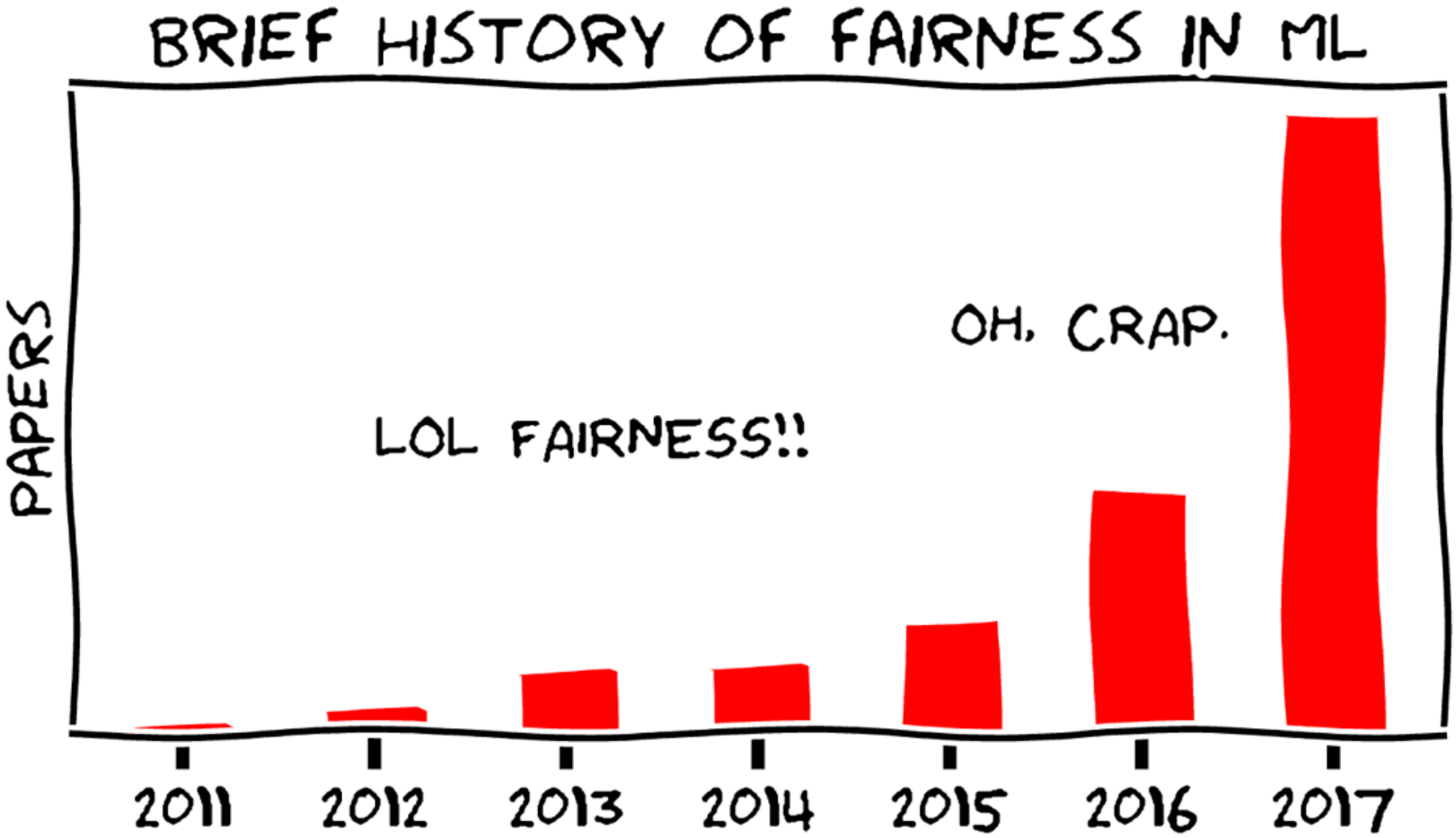
{SEC(2021) 167 final} - {SWD(2021) 84 final} - {SWD(2021) 85 final}

# U.S. Regulation

Under U.S. fair-lending law, lenders can only discriminate for **creditworthiness reasons** (Morse and Pence (2020), Evans (2017)):

- Equal Credit Opportunity Act (ECOA)
- Fair Housing Act (FHA)

# Growing concern



# Reference

Discrimination in lending:

- Bartlett et al. (2021)
- Bhutta and Hizmo (2021)
- Bayer et al. (2018)
- Fuster et al. (2021)

Fairness definitions:

- Verma and Rubin (2018)
- Berk et al. (2021)

# What is an unfair algorithm?

## Unfair algorithm (Definition)

An algorithm is unfair if it systematically places a group of individuals who share a protected attribute (PA), at a systematic disadvantage, e.g., gender, age, residence, ethnic origin, skin color, religion.

The discrimination can be either **direct** or **indirect**.

# Our contributions

1. **Measurement**: Develop inference to see whether the difference between groups is statistically different from zero.

2. **Interpretability**: Identify the variable(s) causing the lack of fairness with a novel tool, the Fairness Partial Dependence Plot (FPDP).

**Does the algorithm comply with the fair lending principle?**

Notations:

- $X \in \mathcal{X}$ : non-protected attributes
- $D \in \mathcal{D} = \{0, 1\}$: protected attribute, with $D = 1$ the protected class
- $f()$: a classification model
- $Y \in \mathcal{Y} = \{0, 1\}$: target variable
- $\hat{Y} = f(X) \in \mathcal{Y} = \{0, 1\}$: predicted outcome

All fairness definitions considered in our paper are based on the joint probability of either $(Y, \hat{Y}, D)$ or $(\hat{Y}, D)$.

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

# Example: Statistical Parity

## Statistical Parity (Definition)

A credit scoring algorithm satisfies Statistical Parity if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class

$$\mathbb{P}(\hat{Y} = 1|D = 1) = \mathbb{P}(\hat{Y} = 1|D = 0).$$

# Example: Conditional Statistical Parity

## Conditional Statistical Parity (Definition)

A credit scoring algorithm satisfies Conditional Statistical Parity if subjects in both protected and unprotected groups have equal probability of being assigned to the positive predicted class, controlling for a set of non-protected attributes

$$\mathbb{P}(\hat{Y} = 1|D = 1, X_c = x_c) = \mathbb{P}(\hat{Y} = 1|D = 0, X_c = x_c).$$

# Fairness test

## Fairness test statistic (definition)

Formally, we denote a fairness test statistic as:

$$F_{H0,i} \equiv h_i(\hat{Y}_j, Y_j, D_j; j = 1, ..., n) = h_i(f(X_j), Y_j, D_j; j = 1, ..., n),$$

where $h_i(.)$ denotes a functional form that depends on the null hypothesis $H0_i$ which is considered, the scoring model $f(.)$ and the sample $\{X_j\}_{j=1}^n$.

Under the null hypothesis $H0_i$ we assume that the test statistic $F_{H0,i}$ has a $\mathcal{F}_i$ distribution, and we denote $d_{1-\alpha}$ the corresponding critical value at $\alpha\%$ significance level.

# Interpretation

If $F_{H0,i} > d_{1-\alpha}$ we reject the null hypothesis of fairness $H0_i$.

Figure 1: Fairness Test traffic lights

# Independence tests

Fairness definitions can be expressed in terms of independence assumptions.

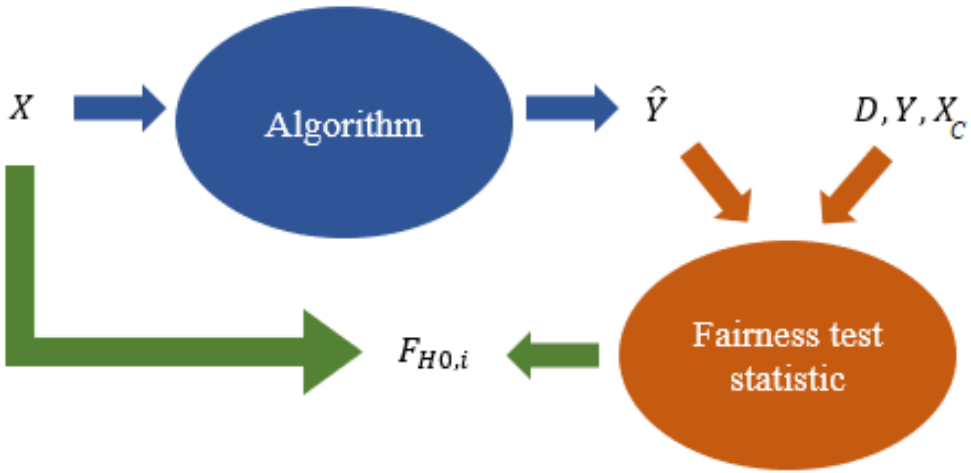
Table 1: Conditional Independence assumption

| Measure | Independence Assumption |
|---|---|
| Statistical Parity | H0: $\hat{Y} \perp\!\!\!\perp D$ |
| Cond. Statistical Parity | H0: $\hat{Y} \perp\!\!\!\perp D \vert X_c$ |
| Equal Odds | H0: $\hat{Y} \perp\!\!\!\perp D \vert Y$ |
| Equal Opportunity | H0: $\hat{Y} \perp\!\!\!\perp D \vert Y = 1$ |
| Predictive Equality | H0: $\hat{Y} \perp\!\!\!\perp D \vert Y = 0$ |

# Objective

Understand the link between the non-protected attributes $X$ and the fairness metric $i \in \{SP, CP, EO, EOP, PE\}$.

Figure 2: Fairness and interpretability



# Fairness Partial Dependence Plot (FPDP)

Assess the marginal effect of a specific feature on a fairness diagnosis test associated to a credit scoring model.

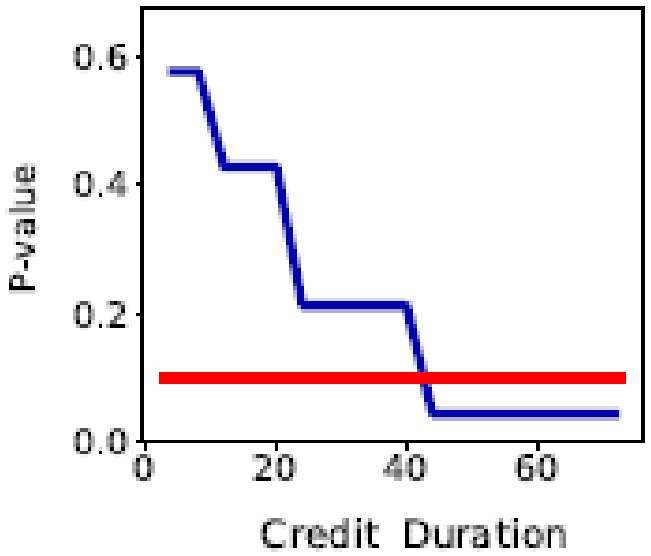
Formally, denote by $X_A$ the feature for which you want to measure the marginal effect, $X_B$ the other features used in the credit scoring model f(.), such that $\hat{Y} = f(X_A, X_B)$. Rewrite the fairness test statistic as:

$$\begin{aligned} F_{H0,i} &\equiv h_i(\hat{Y}_j, Y_j, D_j; j = 1, ..., n) \\ &= h_i(f(X_{A,j}, X_{B,j}), Y_j, D_j; j = 1, ..., n) \\ &= \tilde{h}_i(X_{A,j}, X_{B,j}, Y_j, D_j; j = 1, ..., n), \end{aligned}$$

with $\tilde{h}(.)$ a nonlinear positive function.

# Fairness Partial Dependence Plot (FPDP)

Figure 3: FPDP Example



# Data

German Credit Dataset:

- 1,000 consumer loans
- 20 features including gender the protected attribute
- 690 men and 310 women
- 300 borrowers are in default ($Y = 0$) among which 191 are men and 109 are women

## Figure 4: Feature Distributions

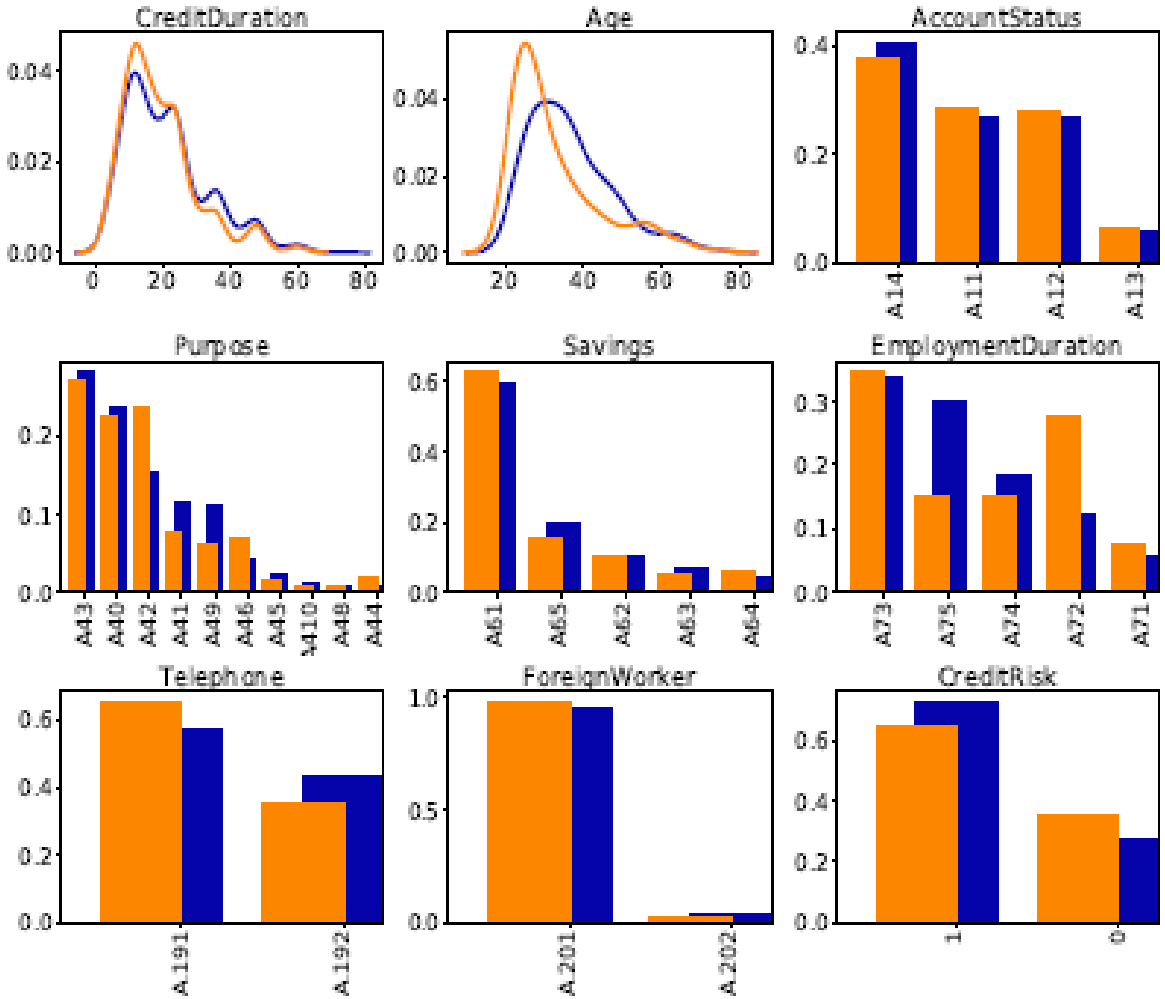
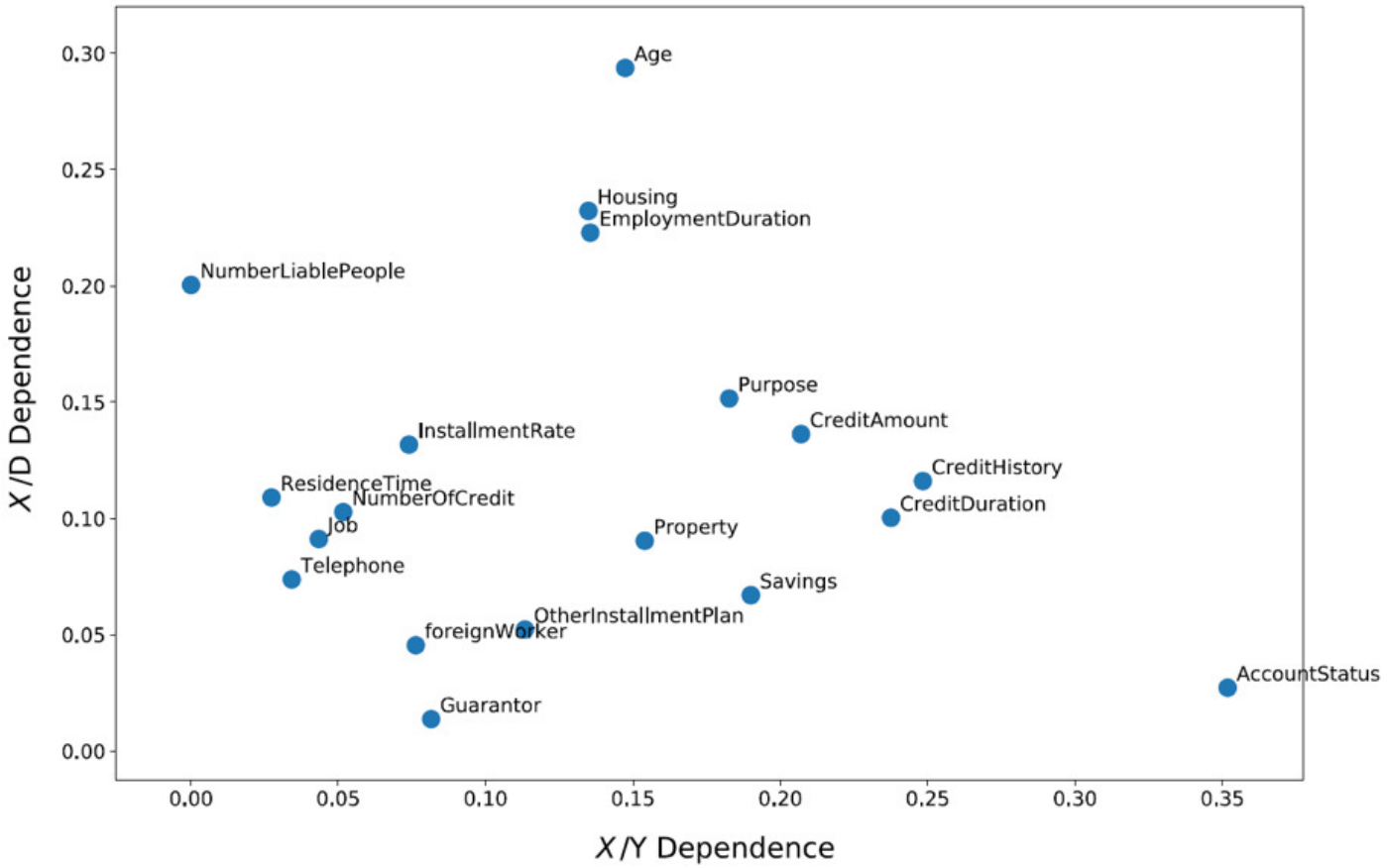


Figure 5: Measures of association between the features, the target variable, and the gender



# Models performance

Table 2: Model performances with and without the protected feature

| | LR | LR (Ridge) | TREE | RF | XGB | SVM | ANN |
|---|---|---|---|---|---|---|---|
| Panel A: Models with gender | | | | | | | |
| 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 |

# Fairness Test

Table 3: Fairness tests for with-models

| | LR | Ridge | TREE | RF | XGB | SVM | ANN |
|---|---|---|---|---|---|---|---|
| Statistical parity | 0.0003* | 0.0001* | 0.0097* | 0.0349* | 0.0004* | 0.0041* | 0.0001* |
| Cond. parity Group 1 | 0.0003* | 0.0001* | 0.0035* | 0.0214* | 0.0004* | 0.0008* | 0.0002* |
| Cond. parity Group 2 | 0.0719 | 0.0781 | 0.4909 | 0.3226 | 0.0417* | 0.3223 | 0.0154* |
| Cond. parity (global) | 0.0003* | 0.0001* | 0.0110* | 0.0434* | 0.0003* | 0.0022* | 0.0001* |
| Equal odds | 0.0185* | 0.0009* | 0.2387 | 0.8220 | 0.0208* | 0.1436 | 0.0084* |
| Equal opportunity | 0.0888 | 0.0105* | 0.3012 | 0.7796 | 0.0206* | 0.1675 | 0.0342* |
| Predictive equality | 0.0242* | 0.0060* | 0.1801 | 0.5733 | 0.1220 | 0.1598 | 0.0242* |

* indicates statistical significance at 5%

# Fairness Test

Table 4: Fairness tests for without-models

| | LR | Ridge | TREE | RF | XGB | SVM | ANN |
|---|---|---|---|---|---|---|---|
| Statistical parity | 0.0734 | 0.1373 | 0.5310 | 0.1206 | 0.1288 | 0.2913 | 0.0159* |
| Cond. parity Group 1 | 0.0989 | 0.0513 | 0.5950 | 0.0966 | 0.2042 | 0.1693 | 0.0086* |
| Cond. parity Group 2 | 0.0866 | 0.4667 | 0.2130 | 0.3226 | 0.0514 | 0.8506 | 0.1445 |
| Cond. parity (global) | 0.0590 | 0.1188 | 0.3998 | 0.1542 | 0.0670 | 0.3821 | 0.0109* |
| Equal odds | 0.6712 | 0.8003 | 0.5645 | 0.9242 | 0.8995 | 0.6753 | 0.3245 |
| Equal opportunity | 0.7746 | 0.9042 | 0.8892 | 0.7796 | 0.7157 | 0.5175 | 0.3917 |
| Predictive equality | 0.3977 | 0.5115 | 0.2890 | 0.7783 | 0.7783 | 0.5451 | 0.2180 |

* indicates statistical significance at 5%

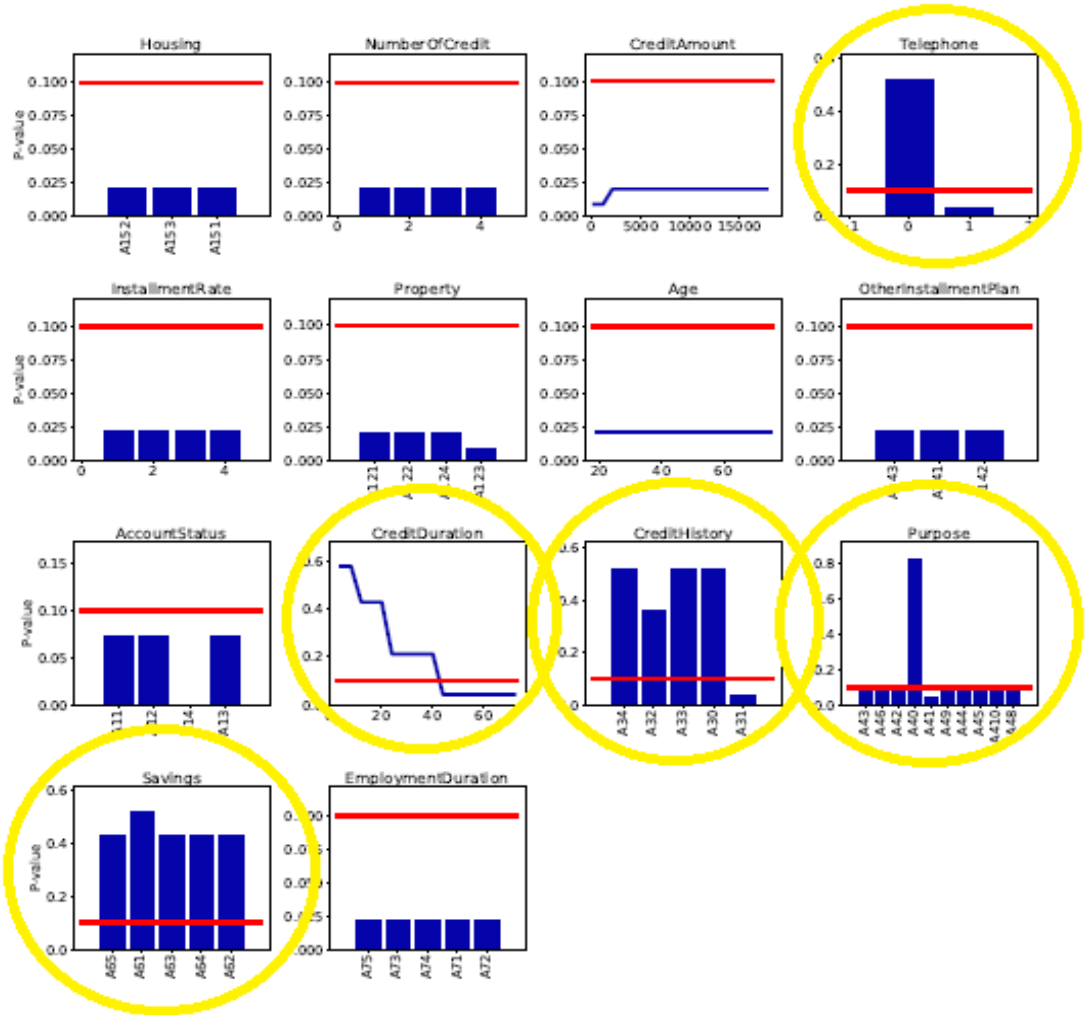
Table 5: Fairness tests for the TREE models

| | TREE | TREE-prime |
|---|---|---|
| Statistical parity | 0.5320 | 0.0216* |
| Cond. parity (global) | 0.3998 | 0.0153* |
| Equal odds | 0.5645 | 0.0363* |
| Equal opportunity | 0.8892 | 0.0101* |
| Predictive equality | 0.2890 | 0.8852 |
| Max. depth of the tree | 1-29 (20) | 1-9 (7) |
| Min. number of individuals required to split a node | 2-9 (2) | 2-59 (56) |
| Min. number of individuals by leaf | 1-19 (5) | 1-59 (18) |
| PCC | 81.5 | 79.0 |
| AUC | 88.6 | 83.9 |

* indicates statistical significance at 5%

# Fairness Interpretability

Figure 6: Fairness PDP for the statistical parity in TREE-prime model



# Associations

Figure 7: Measures of association between features, target variable, and gender, with candidate variables identified for statistical parity test and TREE models

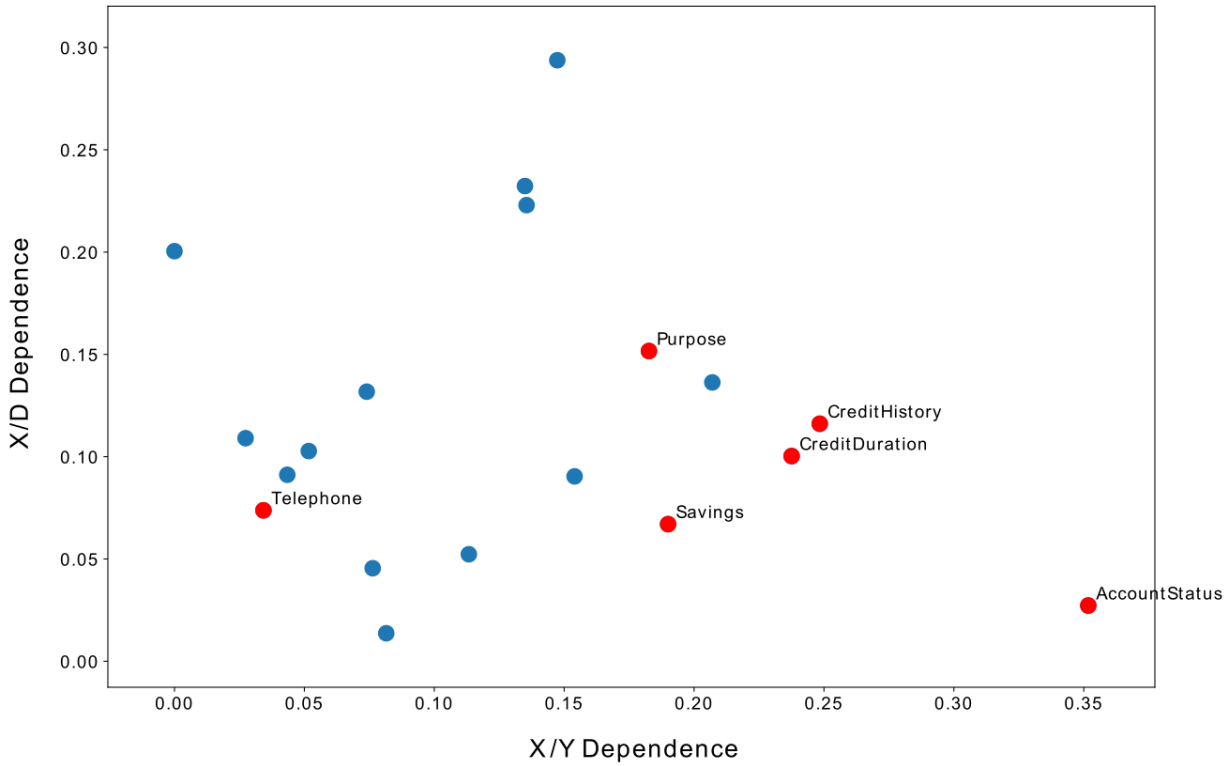


Table 6: Fairness tests for the TREE models

| | TREE-prime | TREE-modif |
|---|---|---|
| Statistical parity | 0.0216* | 0.5195 |
| Cond. parity Group 1 | 0.0552 | 0.7849 |
| Cond. parity Group 2 | 0.0305* | 0.0973 |
| Cond. parity (global) | 0.0153* | 0.2438 |
| Equal odds | 0.0363* | 0.3441 |
| Equal opportunity | 0.0101* | 0.4547 |
| Predictive equality | 0.8852 | 0.2095 |
| PCC | 79.0 | 77.8 |
| AUC | 83.9 | 83.4 |

* indicates statistical significance at 5%

# Conclusion

- Framework to formally check whether their exists a statistical significant difference in terms of rejection rates between protected and unprotected groups ...
- ... and whether this difference is only due to creditworthiness.
- Could be implemented in many other contexts: legal decisions, hiring decisions, claim decisions, etc.