CARMA 2018

# Gender discrimination in algorithmic decision-making

Galina Andreeva[1], Anna Matuszyk[2,3]

[1]*The University of Edinburgh Business School, Galina.Andreeva@ed.ac.uk*

[2]*Stern Business School, New York University, USA;*

[3] *Warsaw School of Economics*

# Outline

- Background: *Gender* prohibition, existing equality legislation
- Empirical analysis of including/excluding *Gender* using a portfolio of car loans
- Impact for lenders (predictive accuracy) and consumers (chances to be accepted/rejected for credit)
- Effect of balancing the training set on *Gender*

Andreeva G., Ansell J., Crook J.N. (2004) Impact of Anti-Discrimination Laws on Credit Scoring. *Journal of Financial Services Marketing,* **9** (1)

Andreeva G., Matuszyk A. (2018). The Law of Equal Opportunities or Unintended Consequences? The Impact of Unisex Risk Assessment in Consumer Credit. Available from SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3212702

# Anti- Discrimination Legislation

- USA
  - Equal Credit Opportunity Act (ECOA, 1974) prohibits characteristics from being used in credit scoring (race, colour, national origin, gender, marital status, religion, receipt of public assistance, or exercise of consumer protection rights). Age has a special status.
- EU
  - Articles 8, 19 of the Treaty of the Functioning of European Union (TFEU);
  - Gender Directive - Council Directive 2004/113/EC of 13 December 2004
  - Proposal for a Council Directive on implementing the principle of equal treatment between persons irrespective of religion or belief, disability, age or sexual orientation, COM(2008) 426 final.

# Economic definitions of discrimination

- Economic theory:
  - Taste-based (subjective discrimination) arises from preferences or prejudice (Becker, 1971);
  - Statistical (objective discrimination) arises from the lack of information necessary to calculate the degree of risk. It is assumed that individuals will behave like the group of which they are members, since there is insufficient information on the individual (Phelps, 1972).

# Legal position

- The Law does not make the distinction between subjective and objective discrimination and is concerned with ‘equal treatment’ or ‘**direct discrimination**’ → certain variables cannot be used as inputs into a model (be it a regression or a machine-learning algorithm).
- But removal of prohibited variables does not automatically create ‘equal outcome’, leading to ‘**indirect discrimination’**, where an apparently neutral criterion would put persons of a particular group (e.g. gender) at a disadvantage compared with other persons, unless that criterion is justified by **a legitimate aim and the means of achieving that aim are appropriate and necessary**.

# Data description

- Portfolio of car loans from a major EU bank from 2003-2010
- Default definition is 2 months (65 days) in arrears
- Women are a minority segment, but with better credit behaviour.

| | Training (80%) | | | Test (20%) | | |
|---|---|---|---|---|---|---|
| | Good | Bad | Total | Good | Bad | Total |
| **Female % by column** | 16746<br>98.70% | 220<br>1.30% | 16966<br>26.71% | 4186<br>98.70% | 55<br>1.30% | 4241<br>26.71% |
| **Male % by column** | 45696<br>98.18% | 847<br>1.82% | 46543<br>73.29% | 11424<br>98.18% | 212<br>1.82% | 11636<br>73.29% |
| **Total % by column** | 62442<br>98.32% | 1067<br>1.68% | 63509 | 15610<br>98.32% | 267<br>1.68% | 15877 |

# Methodology

- Four logistic regression models to predict Probability of Default (PD):

1- Default/Bad; 0- No Default/ Good

1) Model with *Gender* (training sample comprising both men and women)
2) Model without *Gender*
3) Model for men only (training sample consisting of men only)
4) Model for women only (training sample consisting of women only).

- The models are compared from the point of view of predictive accuracy (impact on lenders); and how they affect the chances of men/women to have their credit application rejected (impact for consumers).
- There are 11 final variables selected by significance and predictive accuracy: Marital status, # kids, Income, Time in employment, Profession, Loan duration, Downpayment, Car price, Car age, Phone given, Gender.

# Part of models

| Variable | Attribute/group | Model 1 (with Gender) | Model 2 (no Gender) | Model 3 (male) | Model 4 (female) |
|---|---|---|---|---|---|
| Gender | Female | -0.457 † (0.0867) | | | |
| Net income (ref. Medium income higher) | Low income | -0.4276† (0.1004) | -0.4694† (0.0999) | -0.4565‡ (0.1188) | -0.3478 (0.1945) |
| | Medium income lower | -0.161 (0.1080) | -0.1743 (0.1077) | -0.2504§ (0.1270) | 0.0892 (0.2128) |
| | High income | -0.4551† (0.0945) | -0.4318† (0.0940) | -0.564† (0.1083) | 0.0462 (0.2024) |
| Marital status (ref. Married, No information) | Divorced | 1.9632† (0.1289) | 1.8417† (0.1259) | 2.3574† (0.1599) | 1.0806† (0.2268) |
| | Single | 1.4927† (0.0883) | 1.4617† (0.0873) | 1.685† (0.1041) | 0.8269† (0.1781) |
| | Widowed | 1.1739† (0.2184) | 1.0208† (0.2156) | 1.5121† (0.2843) | 0.3822 (0.3502) |

† - p-value <0.0001, negative sign means females have lower PD

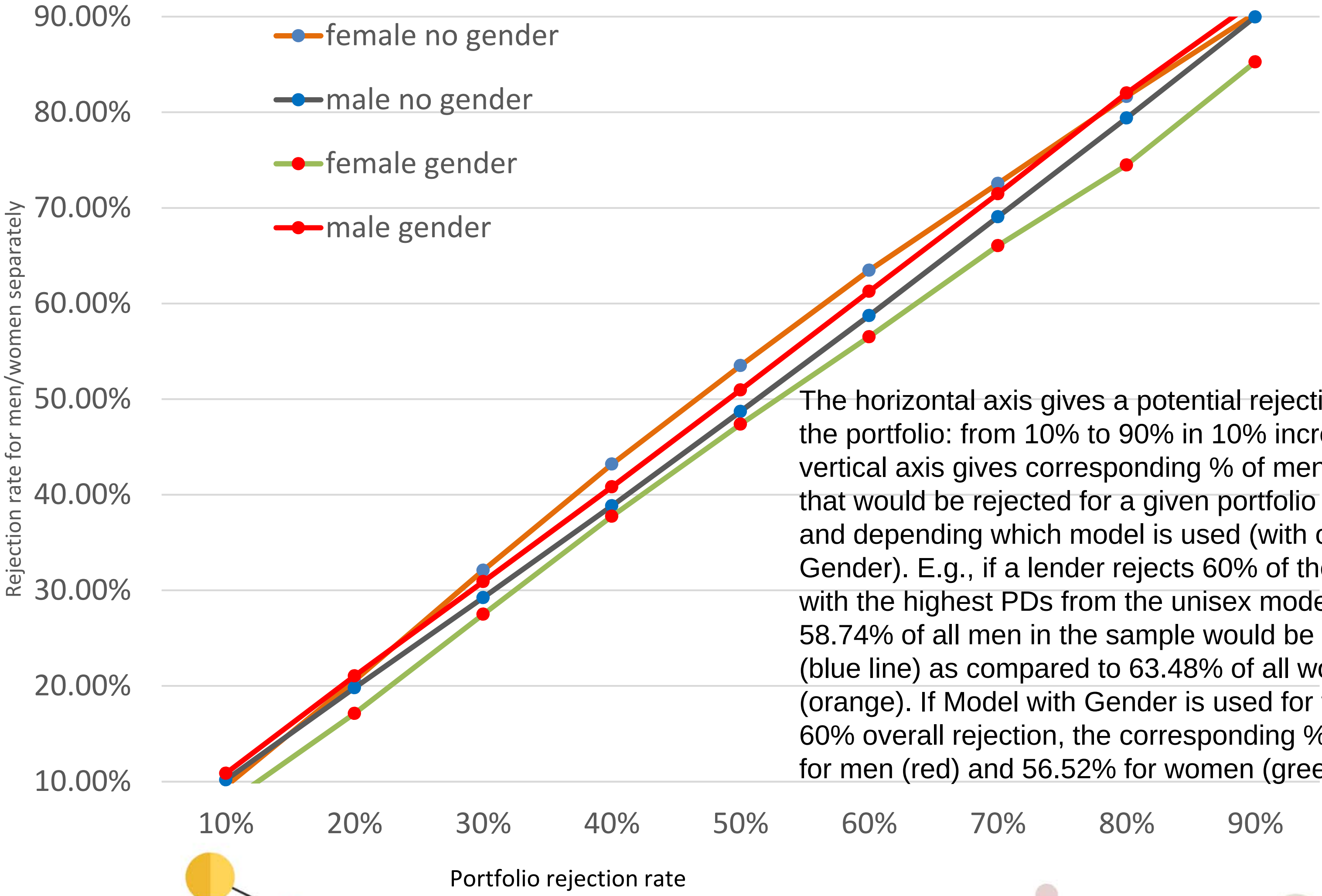
# Predictive accuracy

AUC, test sample (higher values mean better accuracy)

| | Total sample | | | Male only segment | | | Female only segment | | |
|---|---|---|---|---|---|---|---|---|---|
| | Model 1 (G) | Model 2 (noG) | Model 3+4 | Model 1 (G) | Model 2 (noG) | Model 3 | Model 1 (G) | Model 2 (noG) | Model 4 |
| No Balancing | 0.89014 | 0.88984 | 0.89433 | 0.91465 | 0.9139 | 0.9149 | 0.79651 | 0.79434 | 0.80615 |
| Under - sampling of men | 0.89066 | 0.89002 | 0.89397 | 0.91336 | 0.91236 | 0.9141 | 0.8041 | 0.8018 | 0.80615 |
| Over- sampling of women | 0.89022 | 0.8896 | 0.89442 | 0.91377 | 0.91277 | 0.9149 | 0.80056 | 0.79812 | 0.80628 |

Predictive accuracy does not change much for the total sample, which means there is little impact on lenders. But all models predict better for men than for women. This is due to the minority status of women, model training is dominated by patterns typical for men – who are the majority.

# Reject rates by Gender




The horizontal axis gives a potential rejection rate for the portfolio: from 10% to 90% in 10% increments. The vertical axis gives corresponding % of men and women that would be rejected for a given portfolio reject rate and depending which model is used (with or without Gender). E.g., if a lender rejects 60% of the portfolio with the highest PDs from the unisex model (No G), 58.74% of all men in the sample would be rejected (blue line) as compared to 63.48% of all women (orange). If Model with Gender is used for the same 60% overall rejection, the corresponding % are 61.27% for men (red) and 56.52% for women (green).

# Reject rates by Gender with balancing

| Reject rate | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | 100% |
|---|---|---|---|---|---|---|---|---|---|---|
| No Balancing (equivalent to the graph on the previous slide) | | | | | | | | | | |
| F NoG | 9.50 | 20.54 | 32.09 | 43.20 | 53.50 | 63.48 | 72.55 | 81.66 | 90.36 | 100 |
| M NoG | 10.18 | 19.80 | 29.24 | 38.84 | 48.71 | 58.74 | 69.06 | 79.40 | 89.97 | 100 |
| F Gender | 7.64 | 17.12 | 27.49 | 37.75 | 47.37 | 56.52 | 66.05 | 74.49 | 85.26 | 100 |
| M Gender | 10.85 | 21.05 | 30.91 | 40.82 | 50.95 | 61.27 | 71.48 | 82.01 | 91.62 | 100 |
| Under-sampling | | | | | | | | | | |
| F NoG | 9.60 | 20.63 | 31.81 | 42.44 | 52.79 | 63.24 | 72.95 | 82.10 | 90.95 | 100 |
| M NoG | 10.14 | 19.77 | 29.35 | 39.11 | 48.98 | 58.83 | 68.93 | 79.21 | 89.66 | 100 |
| F Gender | 7.66 | 16.69 | 27.28 | 37.44 | 46.95 | 56.45 | 66.66 | 75.36 | 86.30 | 100 |
| M Gender | 10.85 | 21.20 | 30.99 | 40.94 | 51.12 | 61.29 | 71.22 | 81.69 | 91.35 | 100 |
| Over-sampling | | | | | | | | | | |
| F NoG | 9.38 | 20.14 | 31.62 | 42.35 | 52.77 | 62.72 | 71.94 | 81.80 | 90.69 | 100 |
| M NoG | 10.22 | 19.95 | 29.41 | 39.14 | 48.99 | 59.01 | 69.35 | 79.33 | 89.64 | 100 |
| F Gender | 7.50 | 16.67 | 27.00 | 37.14 | 46.80 | 56.19 | 65.88 | 74.32 | 85.66 | 100 |
| M Gender | 10.91 | 21.21 | 31.09 | 41.05 | 51.16 | 61.39 | 71.51 | 82.07 | 91.60 | 100 |

# Conclusions, further research

- Balancing on Gender does not eliminate 'representation bias', since the same set of variables is used, need to re-consider variable selection in addition to sample re-balancing.
- Work in progress includes random forests and neural networks to compare to logistic regression.
- Gender bias is not a simple phenomenon and cannot be eliminated by a simple removal of Gender variable.
- Gender is correlated not only with credit performance, but with other inputs into the model: Income, Occupation, Marital Status (not part of this presentation, see the SSRN paper on slide 2).
- One needs to be careful in order not to give rise to the indirect discrimination.
- A discussion between industry, legislators and academics is needed to arrive to more effective solutions.