Gender discrimination in the age of Big Data: the case of auto loans

Galina Andreeva, *University of Edinburgh, UK,* E-mail: Galina.Andreeva@ed.ac.uk

Anna Matuszyk, *New York University, USA; Warsaw School of Economics, Poland*

1. Introduction

We investigate the consequences of restrictions on information in automated decision-making, using a specific example when the applicant's gender cannot be used as a factor in risk assessment/credit scoring. The results apply generally to situations of algorithmic decisions based on empirical data. Credit scoring is a collection of mathematical and statistical models that predict the probability of a borrower's default, using historic data that may include personal characteristics such as age, income, residential status.

Gender is prohibited by Law from use in decision-making in the majority of developed countries. The prohibition follows from anti-discrimination provisions, e.g. the European Equal Treatment in Goods and Services Directive.

We would like to test empirically the effectiveness of Law in application to automatic decision-making, to highlight potential inconsistencies and inspire further research into better legal solutions. We do it by analyzing a unique proprietary dataset on car loans from an EU bank, which contains gender, other application characteristics and observed credit performance. Our investigation consists in following a standard credit scoring methodology that is used by banks in practice to construct a model based on credit application variables (with and without Gender) and to observe changes in parameter estimates and predictive accuracy.

Table 1. Training and test samples.

	Training			Test		
	Good	Bad	Total	Good	Bad	Total
Female % by column	16746	220	16966	4186	55	4241
	98.70%	1.30%	26.71%	98.70%	1.30%	26.71%
Male % by column	45696	847	46543	11424	212	11636
	98.18%	1.82%	73.29%	98.18%	1.82%	73.29%
Total % by column	62442	1067	63509	15610	267	15877
	98.32%	1.68%		98.32%	1.68%	

2. Methodology

The project follows the standard methodology for building credit scoring models as described in Thomas *et al.* (2002). Logistic regression is the most popular and widely used algorithm in credit scoring and is also used in this paper:

$$logit\ (p_i) = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}_i$$

where p_i is the probability of experiencing default (according to a selected definition) for customer i and $\mathbf{x}_i$ are predictor variables/ characteristics.

The predictors are binned or coarse-classified, which is a standard approach in credit scoring. The resulting coarse-classes are transformed into binary dummy variables.

The estimated probability of default (PD) is used as a 'score', which can be viewed as a summary of credit worthiness. Credit applicants can be ranked on the basis of the score or in other words, according to the level of their attractiveness to lender. The accept /reject decision is achieved by setting a threshold or cut-off: customers with higher probability of default than the cut-off are rejected, whilst those with lower PD are accepted for credit.

Four logistic regression models have been built:

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).

3. Data description

The dataset is a portfolio of car loans coming from a major bank (which chose to remain anonymous) operating in an EU country. Table 1 summarizes the training sample, which is used for the model estimation; and test sample, which is reserved for assessing the model's predictive accuracy. Splitting the data into training and test samples is a standard methodology in credit scoring, here the split is 80% : 20%. 'Bad' are customers who missed two consecutive monthly payments – the definition used by the lender that provided the data.

4. Predictive accuracy

Parameter estimates and model fit measures of the four models are available on request.

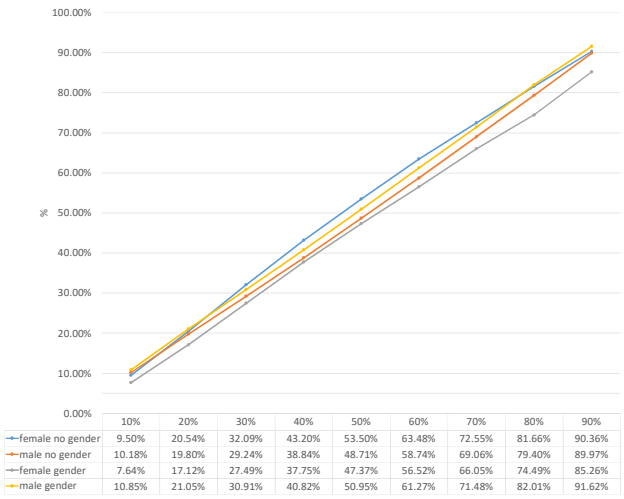
Gender is statistically significant (p-value <0.0001), its removal leads to slightly worse model fit and some changes in parameter estimates. The biggest changes in parameter estimates are observed for 'female' model.

Predictive accuracy is measured by Area under the Roc curve (AUC) and is given in Table 2. Although AUC is higher when separate models are used for both sexes, for males it does not matter much which model is used, whereas for females the uplift is more pronounced. Whilst there is little benefit to women from a simple inclusion of gender into the model, the segmentation does allow capturing unique features of female risk profiles

Table 2. Measures of predictive accuracy training and test sample.

	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
	training sample								
AUC	0.92066	0.92111	0.92381	0.93341	0.93316	0.93330	0.87300	0.87390	0.88596
Sensitivity	0.85473	0.85005	0.85848	0.89138	0.87485	0.87603	0.71364	0.75455	0.79091
1-specificity	0.15502	0.15523	0.15281	0.15938	0.14855	0.14732	0.14314	0.17347	0.16780
	test sample								
AUC	0.89014	0.88984	0.89433	0.91465	0.91390	0.91490	0.79651	0.79434	0.80615
Sensitivity	0.79401	0.78277	0.79026	0.83962	0.82076	0.83019	0.61818	0.63636	0.65455
1-specificity	0.15298	0.15432	0.15112	0.16378	0.15126	0.15100	0.12351	0.16269	0.15504

Figure 1. Impact on rejection by gender when different models are used, % men/women rejected v overall reject rate.



	10%	20%	30%	40%	50%	60%	70%	80%	90%
female no gender	9.50%	20.54%	32.09%	43.20%	53.50%	63.48%	72.55%	81.66%	90.36%
male no gender	10.18%	19.80%	29.24%	38.84%	48.71%	58.74%	69.06%	79.40%	89.97%
female gender	7.64%	17.12%	27.49%	37.75%	47.37%	56.52%	66.05%	74.49%	85.26%
male gender	10.85%	21.05%	30.91%	40.82%	50.95%	61.27%	71.48%	82.01%	91.62%

5. Chances of being accepted/rejected

The ultimate question is how the chances to be accepted for credit are affected. Hand (2012) outlines the following scenarios when talking of potential solutions in achieving discrimination-free credit decisions:

a) Current situation – prohibit the use of gender. One can also consider removing variables with *Gender*, but the question arises, what level of correlation would be acceptable and how many variables would be left for model building.

b) Ensure equal outcome – accept the same proportion of men and women.

Note that only scenario A is legal under the existing regulations, since B requires the use of *Gender* in model-building.

To assess the impact on access to credit, the proportions of men and women have been compared for different cut-off levels that would correspond to a range of rejection/acceptance rates: from 10% to 90% in 10 % increments (Figure 1). E.g., if a lender rejects 60% of the population (sample) and uses PDs from the unisex model (Model 2) as scores, 58.74% of all men in the sample would be rejected as compared to 63.48% of all women, thus the latter segment is not being rewarded for being better credit risks. However, if Model 1 is used for the same cut-off (60% overall rejection), the corresponding percentages become 61.27% for men and 56.52% for women, thus rewarding women for being better credit risks.

Overall, men, being less creditworthy, benefit from unisex model. On the contrary, women would benefit from including *Gender*, since more females would be accepted for credit, However, the removal of *Gender* does not make the reject rates equal for both genders, it almost reverses them, disproportionately punishing women as a more creditworthy class.

6. Conclusion

The results are indicative of the law of unintended consequences. Surely, the main objective of the equality provisions is to protect consumers. Yet, it has been shown that the regulations do not ensure equality of outcome. More creditworthy groups subsidise worse risks, but this subsidy is disproportionate with rejection rates almost reversed.

References

Andreeva, G., Ansell, J., Crook J.N. (2004) Impact of Anti-Discrimination Laws on Credit Scoring, *Journal of Financial Services Marketing,* **9**, 22-33.

Hand, D.J. (2012b) Confusion in scorecard construction – the wrong scores for the right reasons, Presentation at Model Risk in Retail Credit Scoring – Statistical Issues. London. (Available from: http://wwwf.imperial.ac.uk/~abellott/ModelRiskWorkshop.html).

Thomas, L.C., Edelman, D.B., and Crook, J.N. (2002) *Credit scoring and its applications*, Philadelphia: SIAM Publishing.