ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2019)
182, *Part* 4, *pp.* 1287–1311

The law of equal opportunities or unintended consequences?: The effect of unisex risk assessment in consumer credit

Galina Andreeva

University of Edinburgh, UK

and Anna Matuszyk

Warsaw School of Economics, Poland, and New York University, USA

[Received February 2018. Final revision June 2019]

Summary. Gender is prohibited from use in decision making in many countries. This does not necessarily benefit females in situations of automated algorithmic decisions, e.g. when a credit scoring model is used as a decision tool for loan granting. By analysing a unique proprietary data set on car loans from a European bank, the paper shows that gender as a variable in a credit scoring model is statistically significant. Its removal does not alter the predictive accuracy of the model, yet the proportions of accepted women/men depend on whether gender is included. The paper explores the association between predictors in the model with gender, to demonstrate the omitted variable bias and how other variables proxy for gender. It points to inconsistencies of the existing regulations in the context of automated decision making.

Keywords: Algorithmic decision making; Credit scoring; Gender; Statistical discrimination

1. Introduction

This paper investigates the consequences of restrictions on information in automated decision making. It analyses an example from the area of quantitative risk assessment in retail credit, also known as credit scoring, that is used to decide which applicants should be granted credit. Nevertheless, the results can be extended to situations of algorithmic decisions, in general, when predictive algorithms are developed on historic 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 or residential status. Some experts in credit scoring have suggested that large banks, retailers and insurers almost exclusively use automated credit scoring systems to decide whether to accept or reject an application (Anderson, 2017; Thomas *et al.*, 2017). Already in 1997 Hand and Henley wrote

> 'Nowadays it seems that the only organizations which do not use credit scoring approaches are the smaller and/or more personal companies, and those concerned with corporate finance, where statistical methods have been slower to be adopted'

(Hand and Henley (1997), page 531). There is also evidence that even smaller credit institutions, such as microlenders, are adopting this lending technology (Schreiner, 2002).

Address for correspondence: Galina Andreeva, Business School, University of Edinburgh, 29 Buccleuch Place, Edinburgh, EH8 9JS, UK.
E-mail: Galina.Andreeva@ed.ac.uk

This study uses a specific example when gender cannot be included in a credit scoring model as a factor or predictor and illustrates the consequences for a lender, and for male and female credit applicants. Gender is prohibited by law from use in decision making in the majority of developed countries. The prohibition follows from antidiscrimination provisions, e.g. the European Equal Treatment in Goods and Services Directive (European Union Council, 2004). There has been a long debate on how 'equal treatment' should be applied in retail financial services. Initially the Directive included a special clause, allowing the use of gender as a factor in risk assessment in insurance and related financial services, provided that it was justified by 'relevant and accurate actuarial and statistical data' (Article 5(2); European Union Council (2004)). Yet, in March 2011, the European Court of Justice cancelled this provision (European Parliament, 2011). The insurance industry expressed the concern that, since gender is associated with risk, its removal will lead to higher motor premiums for women, who are known to be safer drivers compared with men (Oxera, 2010). The expectation was that the European Court of Justice ruling would make premiums the same irrespective of gender, yet McDonald (2015) showed that the difference in premiums between male and female drivers still remained as a result of 'proxies'—variables which are legal to use and are correlated with both gender and risk, e.g. the customer's profession. This correlation and the resulting insensitivity of the model's predictive power to simple removal of the prohibited variable have been noted before (Andreeva *et al.*, 2004; Hand, 2012a, b). However, there is no clear guidance on how these correlations should be treated and what would be a discrimination-free solution.

In this paper we illustrate empirically the consequences of legal restrictions on information in situations of automated decision making, to highlight potential inconsistencies of the existing regulations and to inspire further research into better solutions. In contrast with McDonald (2015), we do not aim at evaluating the effect of a particular piece of legislation, but we rather wish to illustrate how the principle of 'equal treatment' does not translate into 'equal outcome'.

We do it by analysing a unique proprietary data set on car loans from a European bank. This data set contains gender, other application characteristics and observed credit performance. Our illustration consists in following a standard credit scoring methodology that is used by banks in practice to construct a model based on credit application variables, including gender. We then remove gender (which is statistically significant) and comment on the changes in parameter estimates, which arise because of the omitted variable bias. We find that the predictive accuracy of the model is not affected, but, when we simulate different scenarios for accept–reject decisions by using models with and without gender, we observe differences in proportions of men and women rejected by different models. Furthermore, we apply Bayesian networks (BNs) and multiple correspondence analysis (MCA) to investigate the structure of associations between gender and other predictors in the model. These associations can proxy for gender, and to follow the principle of equal treatment consistently the correlated variables should be excluded from the model in addition to gender. However, when these additional variables are removed, a significant deterioration in predictive accuracy is observed.

The paper contributes to the research and discussion on discrimination in several ways. First, we provide the empirical confirmation of gender's statistical significance in predicting the probability of default in the context of auto loans. Second, we demonstrate empirically the ineffectiveness of existing antidiscriminatory regulations that prohibit the use of gender as a predictor in automated decision making. As noted above, removing gender from the model leaves the predictive accuracy virtually unaffected, which would imply that for lenders there is little difference in terms of which model should be used, since they can achieve a similar level of classification performance. And yet for consumers it matters which model is used as a decision

tool when granting credit. We show that female borrowers have relative higher acceptance rates when the model with gender is used, compared with the model without gender.

Third, we highlight the importance of minority status, i.e., when the protected class constitutes a smaller segment in the overall portfolio or population, it may not be represented appropriately if the distinction between the minority and majority class is not allowed. We use the current analysis as an example to demonstrate what happens to a minority class (women in our case) when this class has better qualities compared with the majority. We believe that the intention of the law should be to protect disadvantaged and/or minority groups. In our case women are better credit risks compared with the majority (men). By prohibiting the use of gender, the law effectively removes the possibility for women to signal their quality and for the lenders to include this signal (gender factor) in the screening process.

Our last (but not least) contribution consists in revealing the associations and dependence structure between all variables in the model, thus accentuating the interconnected nature of this world and the difficulty of eliminating the effect of one single variable or factor.

The relevance of our investigation goes beyond the example of car loans, and we believe that our arguments apply to a wide range of situations where models are built on historic data, and the model captures the relationships of the past, which may include associations of protected groups with perfectly legal variables, such as profession or occupation. The model is then applied to predict the future performance, thus extrapolating historic relationships into the future, and the decisions are made based on those predictions. The timeliness and importance of such investigations are demonstrated by the recent 'Algorithms in decision-making' inquiry of the House of Commons Select Committee, where one of the questions was

> 'the scope for algorithmic decision-making to eliminate, introduce or amplify biases or discrimination, and how any such bias can be detected and overcome . . . '

(Science and Technology Committee, 2017). The submission from the Royal Statistical Society noted the experience from credit scoring as the example worthy of consideration (Royal Statistical Society, 2017).

The remainder of the paper is structured as follows. The next section briefly reviews the relevant economic theories of discrimination and previous empirical research. Section 3 gives an overview of the credit scoring methodology. Subsequently Section 4 describes the data and the results of the empirical analysis. The final section discusses the implications and concludes.

2. Antidiscrimination law, theories of discrimination and empirical research

In general, the law distinguishes between direct and indirect discrimination, although the details may differ across the countries and legislations. Direct discrimination is based on the principle of equal treatment and is unlawful. It arises when a person is treated 'less favourably' because of the prohibited characteristic, such as gender or race. Indirect discrimination arises when a neutral rule or criterion is applied to everyone, but people with protected characteristics are disadvantaged. In the latter case the focus is on equal outcome, and such discrimination can be justified, e.g. by business necessity (European Union Council, 2004).

Economic theory distinguishes between subjective taste-based discrimination (Becker, 1971), which is a consequence of subjective preferences or prejudices, and objective statistical discrimination (Phelps, 1972; Stiglitz, 1993). Statistical discrimination is the consequence of insufficient information that is necessary to estimate the level of risk. Some relevant information (e.g. the intention to repay or the ability to cope with adverse life events) cannot be observed, so, if group membership is correlated or associated with this latent information and both are associated with default, group membership can be used as a signal or a proxy. This

distinction (subjective *versus* statistical) is not reflected in the law. As noted by the Alan Turing Institute (2017), the antidiscrimination legal principles were developed in the era of human judgement and do not cover certain aspects of automated decision making. One such aspect has been pointed out by Hand (2009, 2012a) and is linked to the concept of statistical or group membership discrimination—the law denounces stereotyping based on group membership and calls for the individual approach, yet probabilistic models that underlie the majority of automated decision making rely on summary estimates derived from aggregated data, i.e. a probability cannot be estimated on the basis of a single individual observation.

Previous research has focused mainly on ways of identifying whether discrimination is occurring. One of the most popular approaches is a regression-based methodology where the coefficient for group membership is taken as a measure of discrimination. A vast body of literature is dedicated to this topic, which is not reviewed here, but the interested reader can refer to Andreeva *et al.* (2004) for a summary of this research and to Charles and Guryan (2011) for a detailed discussion of limitations of regression and other approaches to detect discrimination. The most widespread criticism refers to the omitted variable bias, i.e., if group membership is correlated with residuals (unobserved or omitted variables), its coefficient will be biased, so it cannot be interpreted as an indication of discrimination. Given our data for the current study, we cannot explore associations of gender as a group membership indicator with latent constructs that are not observed; however, we can make gender an omitted variable and explore its associations with other predictors in the model.

Another popular strand of literature investigates whether protected characteristics have negative or positive association with risk. Women have been found to be more reliable payers in several studies in a variety of contexts and credit products (e.g. Agarwal *et al*. (2016), Coin (2013), D'Espallier *et al*. (2011) and Do and Paley (2013)). We do not review this topic in detail either, because the objective of this paper is not to prove that women are always better credit risks since this may vary depending on the context and credit product. Instead we investigate empirically the effects of legal restrictions on information in automated decision making and the consequences for a decision maker and the segments that are represented by a prohibited variable (men and women in our case).

Our investigation is linked to recent regulatory inquiries into algorithmic decision making, which inspired research into discrimination arising from machine learning. Fuster *et al*. (2017) showed that more advanced algorithms, such as random forests, increase discrimination, compared with standard logistic regression. This is logical, since higher predictive accuracy means higher discrimination in a statistical sense (i.e. better separation between the risk classes) and, if there are true differences in levels of risk of protected groups, more accurate models will intensify this distinction. Berk *et al.* (2017) quantified this relationship as the 'price of fairness' and showed that, when the predictive accuracy goes up, the 'fairness' will go down. Nevertheless, these references did not analyse the effect of equal treatment on unequal outcome, which is the main focus of this paper.

Several studies (Fair, 1979; Johnson, 2004; Andreeva *et al.*, 2004; Chan and Seow, 2013) have postulated certain implications from the prohibition but have not provided any empirical proof. One implication is that, if a prohibited variable is associated with default, its removal should lead to a reduction in predictive power, which should negatively impact on lenders through increased delinquency and cost of credit. The drop in predictive accuracy may also negatively impact on consumers, if an increased cost of credit is passed on to them.

Another potential implication consists in restricted chances of being accepted for credit, and this can affect protected and unprotected groups. However, the investigation of these implications remains largely speculative in the absence of suitable empirical data which are not

available (except for limited information that is reported for mortgages in the USA under the Home Mortgage Disclosure Act (US Government, 2015). Taylor (2011) noted lack of data as a major obstacle in discrimination research in non-mortgage credit. That is why, having obtained the relevant data, we would like to investigate the predictive value of gender and its effect on access to credit for men and women.

3. Credit scoring methodology

The project follows the standard methodology for building credit scoring models as described in Anderson (2017), Hand and Henley (1997), Siddiqi (2006) and Thomas *et al.* (2017). Our approach consists of mimicking the process of credit model construction in practice, so that observations can be made about the potential effect in a credit granting environment.

Anderson (2017), Hand and Henley (1997), Siddiqi (2006) and Thomas *et al.* (2017) asserted that logistic regression is the most popular and widely used algorithm in credit scoring and it is also used in this paper:

$$\text{logit}(p_i) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i \tag{1}$$

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

The regression is preceded by the predictor variable transformation or coarse classification, which is a standard approach in credit scoring (Jung and Thomas, 2008; Baesens *et al.*, 2009). In the case of continuous variables, the first step is to split a characteristic into intervals, usually between 10 and 20. In the next step, adjacent intervals with similar default rates are merged into larger coarse classes. This enables us to preserve any non-monotonic patterns, to treat outliers and to include missing values that become a separate coarse class. Similarly, for categorical variables, small categories are grouped together to achieve more stable predictions. Finally, the resulting coarse classes are either replaced by weights of evidence or transformed into binary dummy variables (Lin *et al.*, 2012). In this paper we follow the dummy variable approach.

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

In addition to standard measures of model fit, credit scoring models are evaluated in terms of their ability to discriminate between 'good' and 'bad' credit risk. The area under the receiver operating characteristic curve, AUC, is a common measure for the discriminatory power. The receiver operating characteristic curve is obtained by plotting sensitivity against $1 -$ specificity for various cut-off points of the PD, with sensitivity and $1 -$ specificity being defined in the following way:

$$\text{sensitivity} = \text{TPOS}(s)/n_1, \tag{2}$$

$$1 - \text{specificity} = 1 - \text{TNEG}(s)/n_1 \ = \text{FPOS}(s)/n_0, \tag{3}$$

where n_1 and n_0, are the numbers of events or defaults (marked as 1) and non-events (marked as 0) respectively, TPOS is the number of correctly predicted events, FPOS is the number of incorrectly predicted non-events, TNEG is the number of correctly predicted non-events and s is the cut-off that is used to classify predicted probabilities into events or non-events ($p_i \geqslant s$ is classified as an event; $p_i < s$ is classified as a non-event).

In the credit scoring context, sensitivity can be interpreted as a cumulative proportion of defaults or the 'bad' customers with and above a score s (therefore, correctly rejected) and 1 − specificity as a cumulative proportion of non-defaults or the 'good' customers incorrectly rejected (Thomas *et al.*, 2017). The higher values of the AUC would indicate superior models; no discrimination would correspond to an AUC of 0.5. This measure summarizes the ability of the model to rank risk correctly over the whole range of possible scores or cut-offs; therefore it does not require the choice of a specific cut-off.

It was shown by Hanley and McNeil (1982) that conceptually the AUC corresponds to the Wilcoxon or Mann–Whitney statistic, which estimates the probability that a predicted PD of a randomly selected bad account will be higher than or equal to that of a randomly selected good account.

4. Data description and empirical results

The data set is a portfolio of car loans issued from 2003 to 2009 by a major bank (which chose to remain anonymous) operating in a European Union country. Table 1 summarizes the training sample, which is used for the model estimation, and the 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 that was used by the lender that provided the data.

As can be seen from Table 1, females have a lower proportion of bad accounts compared with males. It should also be noted that women are in the minority, constituting slightly more than a quarter of the sample. For this reason and because the event to be modelled constitutes only 1.3% for women, and 1.82% for men, we used a random sampling stratified on gender and good or bad status when dividing the data into training and testing sets. This is an accepted practice in predictive modelling, in particular, of rare events, when it is important that both samples are representative of the population(s) of interest. Kohavi (1995) has shown that a stratified random cross-validation is superior in terms of lower variance and bias compared with a simple random cross-validation. A single split into a training and test set can be viewed as one iteration of a cross-validation with $k = 2$ folds.

Four logistic regression models were built:

(a) a model with gender (a training sample comprising both men and women) (model 1);
(b) a model without gender (model 2);

Table 1. Frequencies and percentages for men and women in the training and test samples

	Results for training sample			*Results for test sample*		
	Good	*Bad*	*Total*	*Good*	*Bad*	*Total*
Number of female customers	16746	220	16966	4186	55	4241
% of female customers	98.70%	1.30%	26.71%	98.70%	1.30%	26.71%
Number of male customers	45696	847	46543	11424	212	11636
% of male customers	98.18%	1.82%	73.29%	98.18%	1.82%	73.29%
Total number of customers	62442	1067	63509	15610	267	15877
Total % of customers	98.32%	1.68%	100%	98.32%	1.68%	100%

(c) a model for men only (a training sample consisting of men only) (model 3);
(d) a model for women only (a training sample consisting of women only) (model 4).

We decided to build segmented models 3 and 4 as opposed to including interactions terms, because this approach is preferred in practice (Banasik *et al*., 1996; Bijak and Thomas, 2012; Thomas *et al*., 2017). This is also in line with our desire to demonstrate the possibility of achieving an equal outcome: with segmented models it is easy to accept or reject equal proportions of men and women. Later in Section 4.2 we explore the associations between all the variables that were used in modelling, including gender.

4.1. Parameter estimates, predictive accuracy and rejection rates

The results from the four models are reported in Table 2. We have used two criteria to include a variable in model 1: the variable must be statistically significant at 0.05; and it must show a high predictive power in explaining the PD. We measured the predictive power by the AUC for each variable separately and selected 12 variables with the highest values. Variables that were selected into model 1 with gender are retained in other models to allow for comparisons of parameter estimates.

As explained in Section 3, we followed the established practice in credit scoring of coarse classification (or categorization) and used dummy variables to represent the final categories. The reference category is selected to be the largest category.

The variables that show significant statistical effects are consistent with the general literature on credit scoring. Shorter loan duration and longer time in employment are associated with the lower PD (Hand, 1998; Thomas *et al*., 2017). Females as well as customers having a job that is typical for women are more creditworthy, similarly to married applicants, and those who provide both commercial and home phone numbers. Having children increases the chances of default.

Considering the features of the car purchased, customers buying cheaper vehicles, with medium capacity engine and having the down payment over 50% of the car price are less likely to default, whereas, in the case of the customers buying older cars, the risk is rising.

When gender is removed (the model without gender), small changes in the parameter estimates can be observed, e.g. the parameter estimates go down for marital status and net income, but increase for loan duration and time in employment. This illustrates the situation of omitted variable bias when the protected characteristic is not included in the model, yet it is associated with the remaining predictors and the dependent variable. The parameter estimates in the model without gender (model 2) still partially reflect its effect.

To understand whether there are differences in the risk profiles of the two sexes, separate models have been fitted to men and women (the last two columns in Table 2—model 3 for men and model 4 for women). Although there are some changes in the estimated parameters between model 3 for men only and models 1 and 2 (with and without gender), the variables remain the same, except for some differences in the magnitude of the parameter estimates. In model 4 for women seven out of 25 categories that were significant in model 1 become insignificant, which may be the result of women being a small segment in the sample. In contrast 3+ kids, which is not significant in model 3 for men, becomes highly significant in model 4 (female). For other categories that remain significant, there is a pronounced change in magnitude; for example the effect of single in marital status for females is almost half of that for males. There are differences of model 4 from models 1, 2 and 3 (with and without gender, men) that are dominated by a larger male segment (around 75%).

Measures of model fit include the Akaike information criterion AIC with lower values indicating better fit, and the pseudo-R^2. Models 1 and 2 (with and without gender) have been

Table 2. Parameter estimates (with standard errors are in parentheses) and model fit statistics for four logistic regression models to predict the PD†

Variable	*Attribute or category*	*% in category*	*Results for model with gender (model 1)*	*Results for model without gender (model 2)*	*Results for model for men only (model 3)*	*Results for model for women only (model 4)*
Intercept			−7.3942‡	−7.5207‡	−7.6844‡	−7.0066‡
			(0.1722)	(0.1708)	(0.2073)	(0.3135)
Gender	Female	26.71	−0.457‡			
			(0.0867)			
Number of children (reference: no kids)	1 kid	23.26	0.19	0.1525	0.267§§	0.1248
			(0.1009)	(0.1000)	(0.1219)	(0.1874)
	2 kids	15.04	0.1918	0.1763	0.2192	0.25
			(0.1302)	(0.1298)	(0.1552)	(0.2447)
	3+ kids	3.12	0.3553	0.3494	0.1584	1.1087§
			(0.2313)	(0.2310)	(0.2950)	(0.3783)
	Missing information	10.87	−0.6816‡	−0.6944‡	−0.7207‡	−0.5017§§
			(0.1254)	(0.1251)	(0.1440)	(0.2618)
Car price (reference: medium price lower)	Cheap	5.28	−1.0987‡	−1.1048‡	−1.2304‡	−0.6552§§
			(0.1326)	(0.1322)	(0.1548)	(0.2718)
	Medium price higher	59.58	0.426§	0.4406‡	0.3752§	0.5243§§
			(0.1099)	(0.1095)	(0.1271)	(0.2277)
	Expensive	15.87	1.1813‡	1.1955‡	1.2106‡	1.0349‡
			(0.1116)	(0.1112)	(0.1286)	(0.2378)
Down payment, % (reference: (35%, 50%])	⩽25%	16.87	1.2702‡	1.2603‡	1.3522‡	0.98‡
			(0.1087)	(0.1085)	(0.1260)	(0.2199)
	(25%, 35%]	8.65	0.7133‡	0.7096‡	0.7704‡	0.515§
			(0.1248)	(0.1246)	(0.1443)	(0.2555)
	>51%	34.49	−1.2147‡	−1.2075‡	−1.2145‡	−1.1765§
			(0.1940)	(0.1941)	(0.2268)	(0.3766)
Car age, years (reference: [0, 2))	2	1.56	1.311‡	1.3197‡	1.3112‡	1.3781‡
			(0.1454)	(0.1448)	(0.1678)	(0.3027)
	[3,4)	3.25	1.8426‡	1.8691‡	1.9397‡	1.4536‡
			(0.1196)	(0.1191)	(0.1362)	(0.2748)
	>4	3.31	2.5302‡	2.5635‡	2.4991‡	2.6423‡
			(0.1348)	(0.1343)	(0.1558)	(0.2835)
Loan duration, months (reference: ⩽36)	(36,60]	17.93	0.7961‡	0.8012‡	0.8606‡	0.586§§
			(0.1181)	(0.1179)	(0.1387)	(0.2320)
	>60	17.08	1.6451‡	1.6554‡	1.6749‡	1.6055‡
			(0.1164)	(0.1162)	(0.1368)	(0.2272)

(*continued*)

Table 2 (*continued*)

Variable	*Attribute or category*	*% in category*	*Results for model with gender (model 1)*	*Results for model without gender (model 2)*	*Results for model for men only (model 3)*	*Results for model for women only (model 4)*
Time in employment, years (reference: ⩾7)	⩽1	11.16	0.7355‡	0.7473‡	0.891‡	0.3584
			(0.1241)	(0.1239)	(0.1477)	(0.2434)
	(1;4]	19.77	1.1924‡	1.2133‡	1.3683‡	0.6768§
			(0.1026)	(0.1022)	(0.1241)	(0.1949)
	(4;7]	25.06	0.7095‡	0.7349‡	0.7505‡	0.6846§
			(0.1107)	(0.1104)	(0.1345)	(0.1995)
Net income (reference: medium income higher)	Low income	26.64	−0.4276‡	−0.4694‡	−0.4565§	−0.3478
			(0.1004)	(0.0999)	(0.1188)	(0.1945)
	Medium income lower	14.90	−0.161	−0.1743	−0.2504§§	0.0892
			(0.1080)	(0.1077)	(0.1270)	(0.2128)
	High income	21.19	−0.4551‡	−0.4318‡	−0.564‡	0.0462
			(0.0945)	(0.0940)	(0.1083)	(0.2024)
Marital status (reference: married, no information)	Divorced	2.99	1.9632‡	1.8417‡	2.3574‡	1.0806‡
			(0.1289)	(0.1259)	(0.1599)	(0.2268)
	Single	16.38	1.4927‡	1.4617‡	1.685‡	0.8269‡
			(0.0883)	(0.0873)	(0.1041)	(0.1781)
	Widowed	2.32	1.1739‡	1.0208‡	1.5121‡	0.3822
			(0.2184)	(0.2156)	(0.2843)	(0.3502)
Car engine capacity, litres (reference: (1.0–1.4])	(1.4−1.6]	10.37	−0.1263	−0.1053	−0.2415	0.1354
			(0.1541)	(0.1538)	(0.1817)	(0.2955)
	>1.6	16.00	0.0925	0.1261	0.0567	0.0592
			(0.1226)	(0.1224)	(0.1409)	(0.2638)
	Missing information	23.39	0.5933‡	0.6119‡	0.5808‡	0.5558§
			(0.1018)	(0.1016)	(0.1196)	(0.1983)
Phone provided (reference: work and home number provided)	Work number provided	13.35	0.1019	0.1148	0.1637	−0.1986
			(0.1073)	(0.1069)	(0.1218)	(0.2447)
	Home number provided	31.20	0.4469‡	0.4472‡	0.4895‡	0.351§§
			(0.0847)	(0.0846)	(0.0984)	(0.1713)
	No phone number provided	15.12	0.3299§	0.3482§	0.3487§§	0.2701
			(0.1114)	(0.1109)	(0.1283)	(0.2351)

(*continued overleaf*)

Table 2 (*continued*)

Variable	*Attribute or category*	*% in category*	*Results for model with gender (model 1)*	*Results for model without gender (model 2)*	*Results for model for men only (model 3)*	*Results for model for women only (model 4)*
Profession or occupation (reference: gender neutral)	Female profession	5.89	−0.5111§§ (0.1938)	−0.6108§ (0.1928)	−0.7843§§ (0.2827)	−0.2068 (0.2653)
	Male profession	13.08	−0.2709§§ (0.1134)	−0.224§§ (0.1129)	−0.2832§§ (0.1246)	−0.2767 (0.3003)
Model fit statistics						
Intercept AIC			10838.202	10838.202	8467.386	2351.084
Intercept and covariates AIC			6976.242	7003.602	5117.254	1833.609
Cox and Snell pseudo-R^2		0.0600	0.0600	0.0595	0.0707	0.0337
Nagelkerke pseudo-R^2		0.3823	0.3823	0.3796	0.4253	0.2606

†The reference category is given in parentheses under the corresponding variable name.
‡ p-value < 0.0001.
§ p-value < 0.005.
§§ p-value < 0.05.

estimated on the same sample; therefore model fit measures can be compared directly, and model 1 with gender demonstrates better fit. This, however, does not translate into superior predictive accuracy (Table 3), where model 2 without gender is slightly better, but the difference is an artefact of random variation (as will be shown in Section 4.3). Models 3 and 4 for men and women have been estimated on different segments; therefore they are not directly comparable, yet a judgement can be made relative to the corresponding measures for models with intercept and intercept plus covariates.

For the AUC that is reported in Table 3 comparisons should also be made within the same sample or segment. For the total sample the column 'Model 3+4 for men and women' refers to estimated PDs that were obtained from segmented models but then combined into a single score to make it comparable with the AUC for models with and without gender (models 1 and 2). Segmented PDs combined do show an uplift in predictive performance, albeit modest. When looking at the predictive accuracy of models applied separately to men or women, for men there is practically no difference between the three PDs, i.e. from model 1, model 2 and model 3. The greatest difference is observed for women. Although there is little difference in predictive accuracy for women from adding gender into model 1, the segmentation does enables unique features of female risk profiles to be captured; hence a modest uplift is observed. The formal tests of difference between AUCs will be reported in Section 4.3.

Yet the ultimate question is what this means for the chances of being accepted for credit. To assess the effect on access to credit, we apply models 1 and 2 (with and without gender) and calculate proportions of men and women who were rejected by each of the two models for different cut-off levels that would correspond to a range of rejection or acceptance rates: from 0.1 to 0.9 in 0.1 increments (Fig. 1). In this part of the analysis we address the concern that was raised by two reviewers who quite rightly noted that in families it is usually men who apply for credit, so it is unfair to claim that all female applicants are disadvantaged since, for many married women, their husbands will apply for credit on their behalf. Therefore, we look at rejection rates for unmarried customers and compare rejected proportions for men and women when they are scored either with model 1 or model 2. Unmarried customers include single, divorced and widowed, and constitute 21.57% of the total (with 8.91% of women and 12.66% of men of the total).

For example, if a lender rejects 60% of the total sample (shown on the vertical axis and the top row under the horizontal axis) and uses PDs from model 1 without gender as scores, 65% of all men in the sample would be rejected compared with 54% of all women (shown on the horizontal axis; first and second rows respectively in the data table). However, if model 1 with gender is used for the same cut-off (60% overall rejection), the corresponding percentages become 71% for men and 45% for women, thus increasing the chances that women will be accepted for credit and rewarding them for being better credit risks.

Overall, men, being less creditworthy, benefit from the model without gender (model 2). In contrast, women would benefit from including gender, since more females would be accepted for credit. It is important to note that the removal of gender does not make the rejection rates equal for both sexes. However, if the objective would be to ensure an equal outcome—accept the same proportion of men and women—this could be achieved with separate models 3 and 4 for men and women.

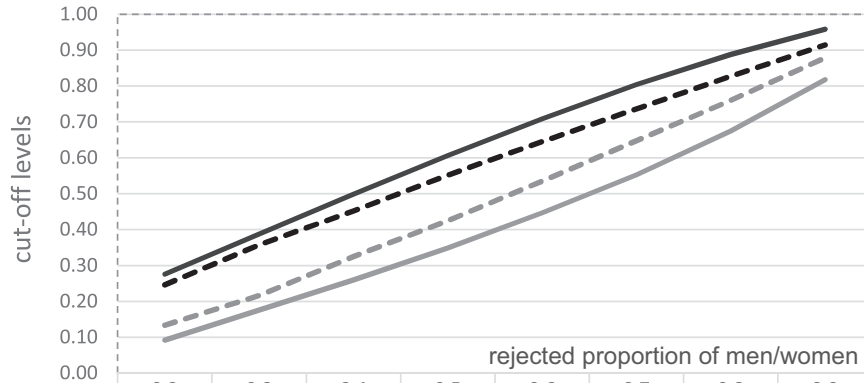
4.2. Interrelationships of gender and other predictors in the models

The results in the previous section demonstrate that, although gender has a statistically significant effect on the PD, its removal from the model does not change the overall predictive accuracy as measured by the AUC (although there are changes in the rejection rates for unmar-

Table 3. Measures of predictive accuracy for the training and test sample: area under the receiver operating characteristic curve, AUC, sensitivity (the proportion of correctly predicted defaults), 1 – specificity (the proportion of incorrectly predicted non-defaults)†

	Results for total sample			*Results for male-only segment*			*Results for female-only segment*		
	Model 1 with gender	*Model 2 without gender*	*Model 3+4 for men and women*	*Model 1 with gender*	*Model 2 without gender*	*Model 3 for men*	*Model 1 with gender*	*Model 2 without gender*	*Model 4 for women*
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

†For sensitivity and 1 – specificity the cut-off was selected at the mean score of the corresponding training sample.



	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
female (M2 no gender)	0.13	0.22	0.33	0.43	0.54	0.65	0.76	0.88
male (M2 no gender)	0.25	0.36	0.45	0.55	0.65	0.74	0.83	0.91
female (M1 with gender)	0.09	0.18	0.26	0.35	0.45	0.55	0.67	0.82
male (M1 with gender)	0.28	0.39	0.50	0.61	0.71	0.80	0.89	0.96

Fig. 1. Effect on rejection by gender when different models are used, considering different proportions of men or women rejected *versus* the overall rejection rate, unmarried customers: the vertical axis and the top row on the horizontal axis show the overall rejection proportion; values in the data table show the rejection proportions for a given group and model

ried women). The apparent lack of change in the predictive accuracy can be explained by the association of gender with other predictors in the model, so, when it is removed, other variables that remain in the model act as a proxy for gender. In this section we explore the association between predictors to understand how this happens.

We use two different techniques: BNs and MCA.

4.2.1. Bayesian networks

BNs are probabilistic models represented by graphs, where random variables are nodes and conditional dependences between them are arrows or arcs. The graph separates the joint (or global) probability distribution of the set of nodes $\mathbf{V} = \{X_1, \ldots, X_v\}$ into a set of local probability distributions: one for each variable. This relies on the Markov property of BNs (Korb and Nicholson, 2010), which implies that a random variable X_i directly depends only on its parents $\text{pa}(X_i)$. In our case we have multinomial (categorized) data; therefore, global and local distributions are given as probability or contingency tables, and

$$P(X_1, \ldots, X_v) = \prod_{i=1}^{v} P\{X_i | \text{pa}(X_i)\}. \tag{4}$$

This formula assumes conditional independence. BNs are often used as causal models even when estimated from observational data (Pearl, 1988). Specifying a causal model of credit default is beyond the scope of this paper. In this section the objective is to identify associations between gender and other variables in the models, and it is sufficient to learn the skeleton of the network or to estimate the undirected essential graph underpinning the network structure. This is the first step in constrained-based algorithms that conduct conditional independence tests to establish relationships that are defined by the Markov property above (Scutari, 2010).

Specifically, to understand the structure of connections in our training sample, we use the 'max-min parents and children' algorithm (Tsamardinos *et al.*, 2006) as implemented in R

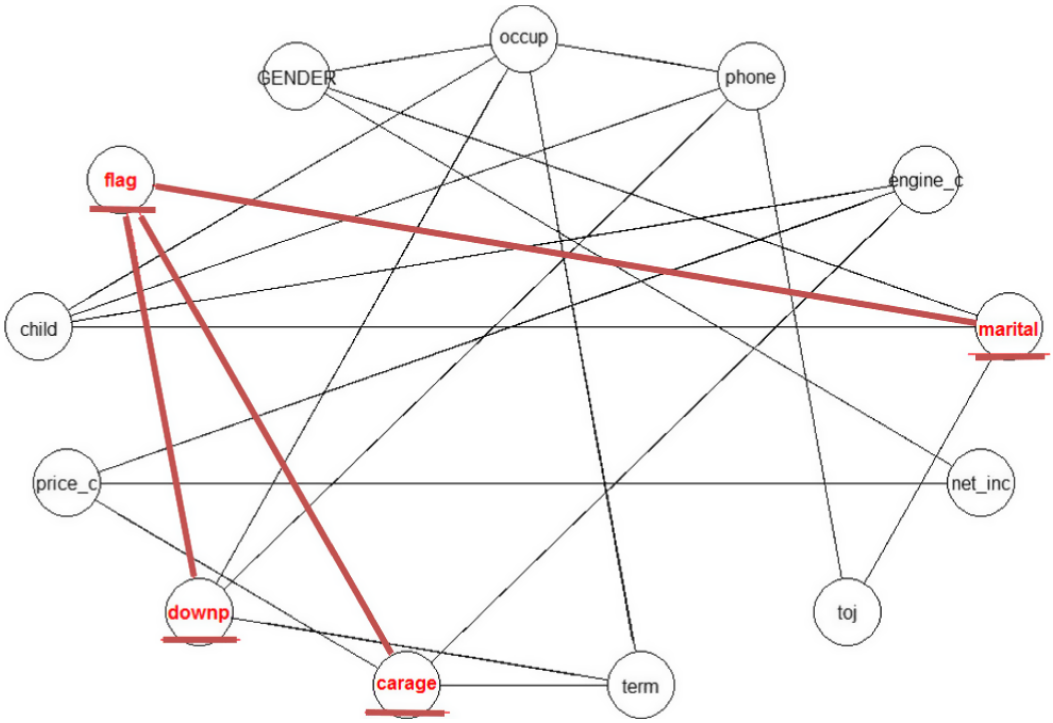


Fig. 2. Markov blanket for flag, default indicator (model 1) (———, direct connections): variables with underlined names are directly related to flag (default indicator); short and long names, occup (occupation or profession), phone (phone number provided), engine_c (car engine capacity), marital (marital status), net_inc (net income), toj (time in employment), term (loan duration), carage (car age), downp (down payment), price_c (car price), child (number of children)

package `bnlearn` (R Development Core Team, 2010; Scutari, 2010). It considers all possible pairs of parents–children and then removes those that are not directly connected. It uses the mutual information between categorical variables (Tsamardinos *et al.*, 2006; R Development Core Team, 2010). The rest of the analysis in this paper has been done by using SAS 9.4 software.

Figs 2–4 show Markov blankets (collections of the node's direct connections) for the variables of interest. Fig. 2 presents the Markov blanket for *flag*, the default indicator, and the dependent variable that is of main interest in credit risk modelling. Fig. 3 shows the Markov blanket for gender, which is of main interest for this study. Figs 2 and 3 demonstrate that gender is not directly connected to flag, but flag is directly related to down payment, car age and marital status. Marital status in its turn is directly related to gender. Gender is also directly linked to occupation and net income. When gender is removed from the model (Fig. 4), the direct link between flag and car age only remains. Therefore, even if gender is not directly related to the outcome, its removal affects the structure of associations between the remaining variables.

4.2.2. Multiple correspondence analysis

Based on the dependence structure that was revealed in the previous section, this section explores the relationship between gender and other predictors with MCA. To make results easier to interpret and present graphically, we have restricted the variables that are used in this section to

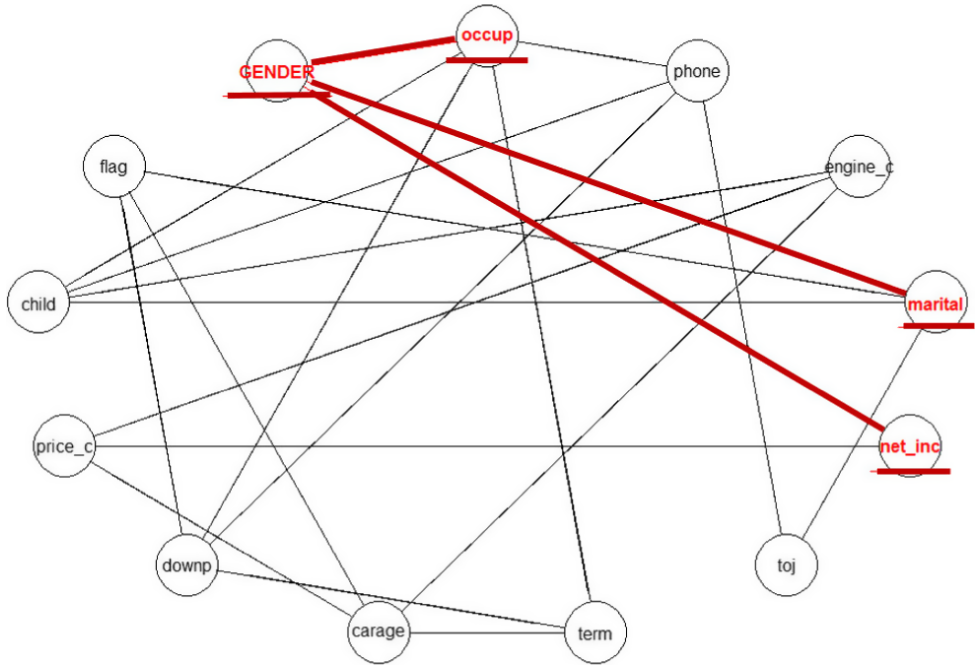


Fig. 3. Markov blanket for gender

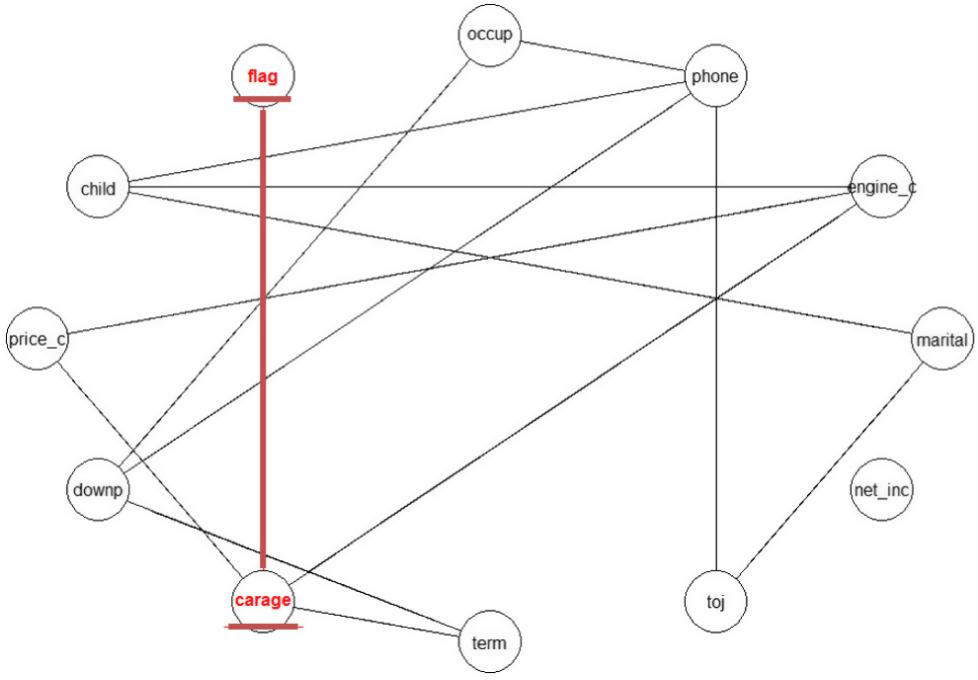


Fig. 4. Markov blanket for flag (the default indicator) with gender removed (model 2)

those that are directly related to gender in the Markov blanket (Fig. 3), because among all the variables that were used not all of them show a direct connection to gender. And it will be very difficult to provide a concise interpretation when the analysis is done on all 13 variables with the corresponding categories. In addition, number of children has also been included given its importance in describing the family status of credit applicants and, therefore, relevance for this study. Number of children is directly related to marital status and occupation, which are in the Markov blanket of gender.

MCA explores the relationship of multiple categorical variables similarly to principal component analysis for numeric variables (Beh and Lombardo, 2014; Greenacre, 2016; Sourial *et al.*, 2010). The result of the MCA is a graphical representation of a contingency table to show associations between the qualitative variables of interest in a low dimensional space. This shows the patterns which cannot be revealed with pairwise analysis. The association or proximity is based on a χ^2-statistic or inertia, which is decomposed into eigenvalues, similarly to how the variance is decomposed in principal component analysis.

The χ^2-statistic summarizes the deviations in a contingency table between the observed frequencies and the frequencies that are expected under the assumption of independence or homogeneity of the categorical variables:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \tag{5}$$

where n_{ij} is the observed value in row r_i and column c_j; and e_{ij} is the corresponding expected value. The inertia is a measure of deviation from homogeneity (or how much variation is contained in the table), which is not dependent on the sample size n:

$$\phi^2 = \chi^2 / n. \tag{6}$$

Following Greenacre (2016), let $\mathbf{N}$ be an $\mathbf{I} \times \mathbf{J}$ non-negative data matrix. The data matrix is converted to the correspondence matrix $\mathbf{P}$, which is a matrix of relative frequencies:

$$\mathbf{P} = \frac{1}{n}\mathbf{N}, \tag{7}$$

where

$$n = \sum_i \sum_j n_{ij} = \mathbf{1}^{\mathrm{T}}\mathbf{N}\mathbf{1}.$$

Row and column marginal proportions are given by

$$\begin{aligned} \mathbf{r} &= \mathbf{P}\mathbf{1}, \\ \mathbf{c} &= \mathbf{P}^{\mathrm{T}}\mathbf{1} \end{aligned} \tag{8}$$

and diagonal matrices of row and column marginal proportions

$$\begin{aligned} \mathbf{D}_r &= \mathrm{diag}(\mathbf{r}), \\ \mathbf{D}_c &= \mathrm{diag}(\mathbf{c}). \end{aligned} \tag{9}$$

Row profiles are contained in

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{P}, \tag{10}$$

where the elements of each row sum to 1. Each (i, j) element of $\mathbf{R}$ is the observed probability of being in column j provided that it is in row i.

Column profiles are contained in

$$\mathbf{C} = \mathbf{D}_c^{-1}\mathbf{P}. \tag{11}$$

The correspondence analysis algorithm computes the co-ordinates based on a generalized singular value decomposition of $\mathbf{P}$:

$$\mathbf{P} = \mathbf{A}\mathbf{D}_\alpha\mathbf{B}^{\mathrm{T}}$$

where $\mathbf{A}$ is the rectangular matrix of left generalized singular vectors, $\mathbf{B}$ is the rectangular matrix of right generalized singular vectors and $\mathbf{D}_\alpha$ is the diagonal matrix of singular values, so that $\alpha_1 > \alpha_2 > \alpha_3 > \ldots$, and

$$\mathbf{A}^{\mathrm{T}}\mathbf{D}_r^{-1}\mathbf{A} = \mathbf{B}^{\mathrm{T}}\mathbf{D}_c^{-1}\mathbf{B} = \mathbf{I}. \tag{12}$$

The principal inertias are given by

$$\lambda_k = \alpha_k^2, \qquad k = 1, 2, \ldots, K,$$

where $K = \min\{I - 1, J - 1\}$.

In MCA

$$\mathbf{P} = \mathbf{B}\mathbf{D}_\alpha^2\mathbf{B}^{\mathrm{T}}, \tag{13}$$

because in this paper we use MCA based on the Burt matrix, which gives all two-way cross-tabulations of the Q categorical variables with a number of categories $J = \Sigma_q J_q$. The number of principal inertias or dimensions in MCA is at most $J - Q$. For more details on the theory of correspondence analysis and MCA, refer to Greenacre (2016).

Table 4 shows the decomposition of total inertia with corresponding statistics, with the singular value indicating the relative contribution of each dimension to an explanation of the inertia, or proportion of variation. The principal inertia is an indicator of how much of the variation in the original data is retained in the corresponding dimension. Table 4 indicates that all variation (or 100% of total inertia) in the Burt matrix can be represented by 13 dimensions (the total number of categories—18—minus the total number of variables used—5). The contribution of

Table 4. Inertia and χ^2-decomposition

Singular value	*Principal inertia*	χ^2	*%*	*Cumulative %*
0.5533	0.2844	93455	10.94	10.94
0.4923	0.2424	79651	9.32	20.26
0.4655	0.2167	71226	8.34	28.60
0.4594	0.2111	69369	8.12	36.72
0.4530	0.2052	67436	7.89	44.61
0.4493	0.2019	66355	7.77	52.38
0.4467	0.1995	65570	7.67	60.05
0.4456	0.1986	65260	7.64	67.69
0.4434	0.1966	64595	7.56	75.25
0.4385	0.1923	63201	7.40	82.65
0.4218	0.1779	58464	6.84	89.49
0.3846	0.1479	48614	5.69	95.18
0.3543	0.1255	41259	4.82	100.00
—	2.6000	854455	100.00	—

each dimension in total inertia (as shown in the column '%' of Table 4) ranges as shown from approximately 11 to 5, with no dimension encompassing the major part of the data.

Nevertheless, there are some high contributions of categories in selected dimensions as shown in Table 5, which presents three of the largest dimensions (detailed results on other dimensions are available on request). Here the main variable of interest, gender, only shows substantial loading to the first two dimensions: female is positively associated with dimensions 1 and 2, with male showing negative association; therefore, here the differences between sexes are most apparent. Dimension 3 shows high loadings or associations of income.

The interpretation of the results is complemented by Figs 5–7 where the judgement depends on the extent to which the categories are relatively close to each other. In Fig. 5, the first dimension on the horizontal axis reflects a split of the clients on marital status and number of children. In Fig. 5 divorced D, widowed W and single S are on the positive side of dimension 1, with married M appearing on the opposite (negative) side, although M is relatively close to the origin or centroid of the data and, therefore, does not contribute much to the association, whereas in the case of number of children the three categories 1, 2 or 3+ kids are on the negative side of dimension 1, opposite mis kids (missing information) and no kids (these two are close to each other and the origin). Dimension 2 (Fig. 5) separates gender categories, putting them on opposite sites, namely female in the upper–positive part and male below. Furthermore, dimension 2 separates female occupation (in the top right-hand corner) from two remaining categories: neutral and male occupation. Dimension 3 (Figs 6 and 7) clearly reflects a division of the income categories, namely low (Low inc) and lower middle income (Mid inc1) *versus* two other categories, i.e. higher middle income (Mid inc2) and high income (High inc).

Categories that are related appear in the same quadrants, although the distance is approximate. For example, the top right-hand quadrant of Fig. 5 shows that the following variable categories are associated: female, female occupation, widowed and divorced. Similarly, in the

Table 5. Contribution of each category to each dimension

Variable	*Category*	*Results for dimension 1*	*Results for dimension 2*	*Results for dimension 3*
Gender	Female	0.8232	0.9901	−0.0875
	Male	−0.3001	−0.3609	0.0319
Number of children	1 kid	−0.6188	0.7525	−0.2837
	2 kids	−0.971	0.3206	−0.306
	3+ kids (3 or more children)	−1.2519	−0.1746	0.5808
	mis kids (missing information)	0.1841	−0.1729	1.8177
	no kids (no children)	0.6476	−0.4171	−0.2174
Occupation	female occ (occupations with female majority)	0.8005	1.6965	0.6333
	male occ (occupations with male majority)	0.0249	−1.0634	0.1797
	neut occ (gender neutral occupations)	−0.0622	0.0484	−0.075
Marital status	D (divorced)	1.084	2.3252	1.4982
	M (married)	−0.393	0.0406	−0.1278
	S (single)	1.4648	−0.789	0.3786
	W (widowed)	1.532	1.2036	−0.2901
Income	Low inc (low income)	0.3809	−0.0413	−1.0002
	Mid inc1 (lower middle income)	0.1369	0.0816	−0.4144
	Mid inc2 (higher middle income)	−0.0633	−0.0327	0.3604
	High inc (high income)	−0.4638	0.0519	0.9153

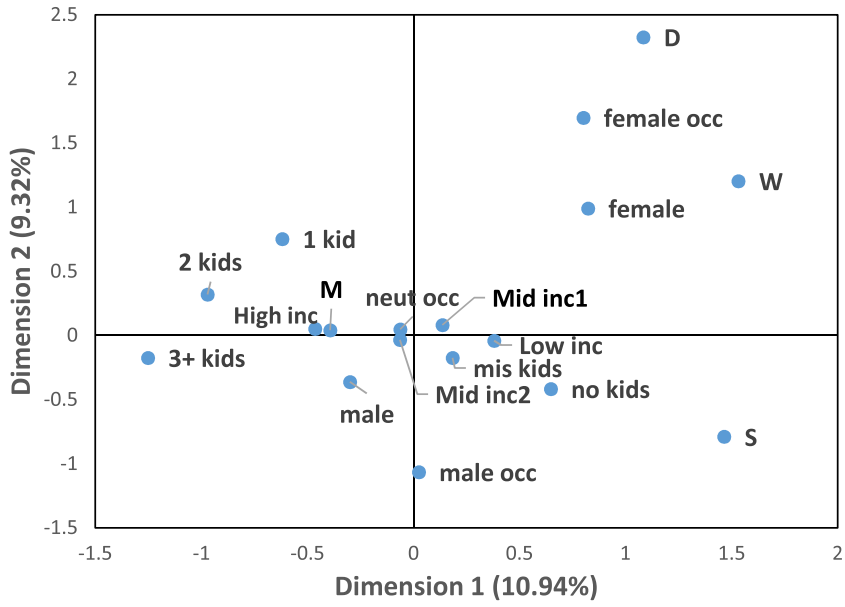


Fig. 5. MCA: position of categories against dimension 1 and dimension 2 (the distance is approximate and represents the relative position of categories from the origin or centroid; full category names are given in Table 5)

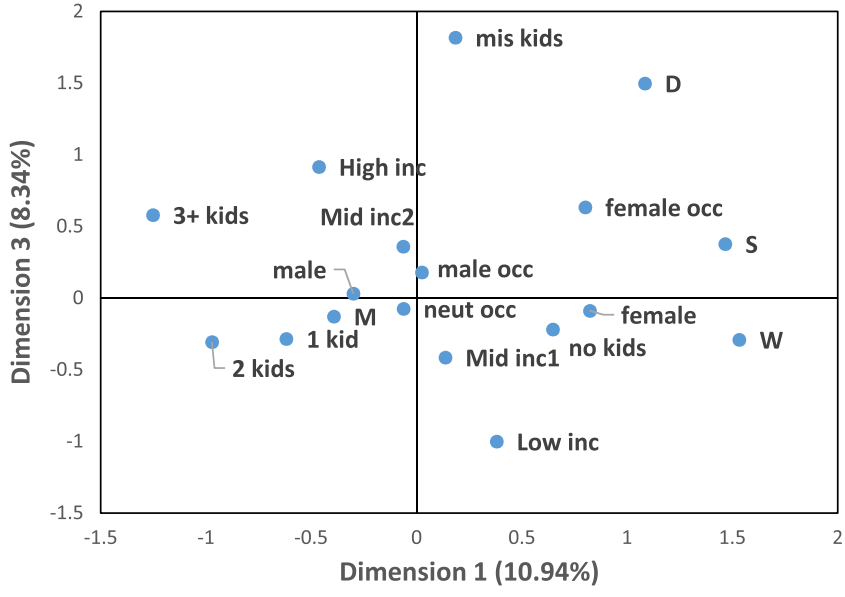


Fig. 6. MCA: dimension 1 and dimension 3 (the distance is approximate and represents the relative position of categories from the origin; full category names are given in Table 5)

bottom left-hand quadrant male and 3+ kids appear together. The other categories close to male, e.g. higher middle income (Mid inc2) or neutral occupation (neut occ) are almost at the origin, so there is little association. In Fig. 7 male appears at the origin of dimension 3 (the dimension that is associated with net income), female is not far from the origin but still appears on the same side as low income (Low inc) and lower middle income (Mid inc1). One can say that

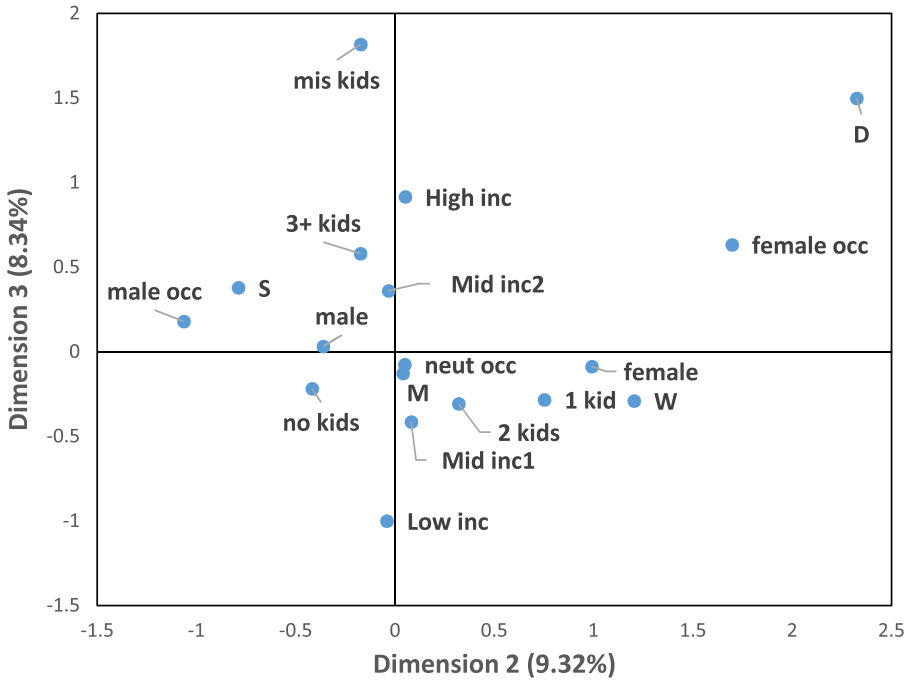


Fig. 7. MCA: dimension 2 and dimension 3 (the distance is approximate and represents the relative position of categories from the origin; full category names are given in Table 5)

female borrowers tend to be widowed or divorced and with lower income compared with males. This makes the effect of gender removal from risk assessment even more socially significant; it is not simply the females in general who are disadvantaged by this policy, but socially vulnerable groups within the female segment

4.3. Comparison of predictive accuracy

In this section we follow another suggestion from one of the referees and investigate whether the predictive accuracy changes when variables that can proxy for gender are removed from the model. In fact, such a situation may arise if regulators require not only prohibited variables to be removed from the statistical models, but also variables that are connected to the prohibited variables. A possible choice is to use variables in the Markov blanket for gender, i.e. those directly connected to it, as identified in Section 4.2.1: marital status; net income; occupation. We label these as models 1(a) or 2(a). We also subjectively add to the list of removed variables number of children because of its connection with marital status, and label them models 1(b) or 2(b). The full list of models that were developed and tested is as follows:

(a) model 1—all variables, including gender;
(b) model 2—no gender;
(c) model 1(a)—no marital status, net income, occupation;
(d) model 2(a)—no gender, marital status, net income, occupation;
(e) model 1(b)—no marital status, net income, occupation, number of children;
(f) model 2(b)—no gender, marital status, net income, occupation, number of children;
(g) model 3+4—combined PDs from model 3 (male) and model 4 (female); see Section 4.2.

Table 6. Results of the tests of difference in predictive accuracy of models with different variables†

Model	*AUC*	*Results for the following models:*						
		Model 1	*Model 2*	*Model 1(a)*	*Model 2(a)*	*Model 1(b)*	*Model 2(b)*	*Model 3+4*
Training sample								
1	0.9207	—	0.3342	<0.0001	<0.0001	<0.0001	<0.0001	0.0041
2	0.9211	0.3342	—	<0.0001	<0.0001	<0.0001	<0.0001	0.0014
1(a)	0.9117	<0.0001	<0.0001	—	0.0391	0.4946	0.1763	<0.0001
2(a)	0.9119	<0.0001	<0.0001	0.0391	—	0.0273	0.3525	<0.0001
1(b)	0.9104	<0.0001	<0.0001	0.4946	0.0273	—	0.0599	<0.0001
2(b)	0.9107	<0.0001	<0.0001	0.1763	0.3525	0.0599	—	<0.0001
3+4	0.9238	0.0041	0.0014	<0.0001	<0.0001	<0.0001	<0.0001	—
Test sample								
1	0.8901	—	0.7704	0.0074	0.0011	0.0093	0.0016	0.2034
2	0.8898	0.7704	—	0.0127	0.0020	0.0103	0.0015	0.1822
1(a)	0.8809	0.0074	0.0127	—	0.1150	0.6428	0.2517	0.1632
2(a)	0.8813	0.0011	0.0020	0.1150	—	0.0994	0.6378	0.0464
1(b)	0.8785	0.0093	0.0103	0.6428	0.0994	—	0.1255	0.1521
2(b)	0.8789	0.0016	0.0015	0.2517	0.6378	0.1255	—	0.0392
3+4	0.8943	0.2034	0.1822	0.1632	0.0464	0.1521	0.0392	—

†Values in the third–ninth columns are *p*-values for a pairwise test of difference against the null hypothesis of no difference: Pr $> \chi^2$ test statistic.

The receiver operating characteristic curves that we are about to compare are developed and applied to the same training and test samples, so they and their corresponding areas (AUC) are correlated. Therefore, we use the test for correlated curves that was proposed by DeLong *et al.* (1988), which tests that the selected pair of AUCs are significantly different against the null hypothesis of no difference.

As noted in Section 4.2, the removal of gender (model 1 *versus* model 2) does not affect the predictive accuracy and this is confirmed by high *p*-values greater than 0.05 in Table 6. However, model 1 significantly differs from all the other models, although it should be noted that the difference with segmented model 3+4 becomes insignificant in the test sample. So the removal of variables that are correlated with gender affects the predictive accuracy of credit scoring models. If gender is then added to the model with correlated variables removed (model 1(a) *versus* model 2(a)), the result is insignificant in the test sample. The same applies to models without number of children (model 2(a) *versus* 2(b)). Therefore, we cannot conclude that gender on its own is a powerful predictor, despite its being a statistically significant variable in model 1.

This is one of the questions that can be explored further: there seems to be a weak effect on predictive accuracy of a single variable or credit scoring models may not be sensitive to the removal of just one variable, yet the effect becomes significant for the combination of variables (although it may depend on the exact combination of variables in the model).

5. Concluding remarks

This paper has explored the concept of equal treatment (which is the fundamental principle of antidiscrimination regulations) in application to statistical or algorithmic discrimination in automated decision making. We have used the case of retail credit risk screening and examined equal treatment of men and women, which translates into the prohibition to use gender in credit

scoring models. The potential effect of the prohibition is considered for lenders (any effects on predictive power of models, or their ability to distinguish between good and bad borrowers), and consumers (their chances of being accepted or rejected for loans). Our illustration of potential impact is based on a data set of applications for auto loans, where women form a minority and constitute about a quarter of the portfolio. Following a standard credit scoring methodology, four models were built initially: model 1 with and model 2 without gender as a dummy variable, and two separate models for men and women (model 3 and 4 respectively).

Gender is statistically significant as a dummy variable, yet its removal does not have a pronounced effect on the predictive power of the model. This is due to the correlation or association of gender with other characteristics that are not legally restricted and remain in the model. If predictive accuracy is unaffected, lenders can maintain similar levels of bad debt for a given acceptance level irrespective of which model is used; therefore, there is no effect on lenders. It does not mean though that they will accept the same applicants when using different models.

The paper then concentrates on unmarried applicants to investigate the effect for consumers and shows that female applicants, having lower default rates in the past, benefit from model 1 with gender, which gives them extra points for being good risks and, therefore, increases their chances of being accepted for credit. Their rejection rates are lower compared with those for men. When applying unisex model 2 (without gender), the chances of being accepted for credit decrease for women but increase for men, although, in general, females still exhibit lower rejection rates compared with men. So the prohibition does not lead to equality in the outcome. It should be noted though that, if the desire is to reject the same percentage of men and women, this can easily be achieved when both groups are treated separately (models 3 and 4) and the same proportions of both sexes that show better ratings on the corresponding scores are selected for credit.

The results are indicative of the law of unintended consequences. Surely, the main objective of the equality provisions is to protect disadvantaged groups of consumers. Yet, it has been shown that the regulations do not ensure equality of outcome. More creditworthy groups subsidize worse risks, and justification for this subsidy could be researched further.

Another potential pitfall for achieving equality is situations when protected groups constitute a minority and, therefore, are not equally represented in the data and, one can argue, not equally treated, as the result of this. In our case the unisex model 2 is dominated by a majority group, and minority unique features are not captured by a 'politically correct' model (that does not contain a prohibited variable).

One more controversy that we would like to highlight is the difficulty of removing the effect of prohibited variables. We have shown that gender is significantly associated with other 'legal' variables; therefore, even when removed, the prohibited variable still partially influences the model through associations. This explains the relative insensitivity of predictive accuracy to gender removal and the remaining difference in rejection rates between men and women. However, there is no clear legal guidance on potential solutions in such situations, i.e. what level of association would be (un)acceptable?

In general, the paper has highlighted certain inconsistencies in the existing framework when it is being applied to statistical discrimination. These highlights are particularly timely given the increasing applications of machine learning algorithms in decision making. Nevertheless, this study is not without limitations. An obvious limitation is that we investigate the accept or reject decision by using a portfolio of accepted loans. This is, however, a well-known problem in credit scoring: the models are developed on accepted loans, since for rejected loans the outcome variable is not observed. There is a methodology to correct for a potential sample selection bias called 'reject inference', but Crook and Banasik (2004) have investigated a variety of such approaches on a rarely available unbiased population and concluded that reject inference was

not effective. Therefore, we believe that our investigation of predictive accuracy is still valid. As for the effect on the consumers, further research could investigate the overall population of applicants, if rejected applications become available. Yet accepted customers form part of this population, so one can argue that our results hold for at least part of the overall population.

Another limitation consists in the fact that the paper has explored only a limited number of variables. Although these are typical variables as stated by the credit scoring literature (Hand, 1998; Thomas *et al.*, 2017), they are by no means exhaustive in terms of other variables that can be used in different portfolios, countries and credit products. Our study can be viewed as an exploratory illustration highlighting the issues that would warrant further research.

Further investigation of other protected characteristics across a wider range of portfolios and countries would contribute towards better understanding of discrimination mechanisms and finding fairer solutions. Other areas of decision making, such as fraud detection, can also be investigated.

Acknowledgements

The authors are extremely grateful to the data provider and two referees for their diligent approach and helpful comments that have improved the paper substantially.

References

Agarwal, S., He, J., Sing, T. F. and Zhang, J. (2016) Gender gap in personal bankruptcy risks: empirical evidence from Singapore. *Rev. Finan.*, **22**, 813–847.

Alan Turing Institute (2017) Written evidence submitted by The Alan Turing Institute (ALG0073) to 'Algorithms in decision-making' inquiry. Alan Turing Institute, London. (Available from `http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/written/69165.html`.)

Anderson, R. (2017) *The Credit Scoring Toolkit*. Oxford: Oxford University Press.

Andreeva, G., Ansell, J. and Crook, J. N. (2004) Impact of anti-discrimination laws on credit scoring. *J. Finan. Serv. Marktng*, **9**, 22–33.

Baesens, B., Mues, C., Martens, D. and Vanthienen, J. (2009) 50 years of data mining and OR: upcoming trends and challenges. *J. Oper. Res. Soc.*, **60**, S16–S23.

Banasik, J. L., Crook, J. N. and Thomas, L. C. (1996) Does scoring subpopulations make a difference? *Int. Rev. Retl Distribn Consum. Res.*, **6**, 180–195.

Becker, G. S. (1971) *The Economics of Discrimination*. Chicago: University of Chicago Press.

Beh, E. J. and Lombardo, R. (2014) *Correspondence Analysis: Theory, Practice and New Strategies*. Chichester: Wiley.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M. J., Morgenstern, J., Neel, S. and Roth, A. (2017) A convex framework for fair regression. *arXiv Preprint*. (Available from `https://arxiv.org/abs/1706.02409`.)

Bijak, K. and Thomas, L. C. (2012) Does segmentation always improve model performance in credit scoring? *Exprt Syst. Appl.*, **39**, 2433–2442.

Chan, W. L. and Seow, H. (2013) Legally scored. *J. Finan. Reguln Complnce*, **21**, 39–50.

Charles, K. K. and Guryan, J. (2011) Studying discrimination: fundamental challenges and recent progress. *Working Paper 17156*. National Bureau of Economic Research, Cambridge.

Coin, D. (2013) Are female entrepreneurs better payers than men? *Questioni di Economia e Finanza, Banca D'Italia*, no. 186. Bank of Italy, Rome. (Available from `http://www.bancaditalia.it/pubblicazioni/econo/quest_ecofin_2/qef186/QEF_186.pdf`.)

Crook, J. and Banasik, J. (2004) Does reject inference really improve the performance of application scoring models? *J. Bankng Finan.*, **128**, 857–874.

DeLong, E. R., DeLong, D. M. and Clarke-Pearson, D. L. (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, **44**, 837–845.

D'Espallier, B., Guérin, I. and Mersland, R. (2011) Women and repayment in microfinance. *Wrld Devlpmnt*, **39**, 758–772.

Do, C. and Paley, I. (2013) Does gender affect mortgage choice?: Evidence from the US. *Femin. Econ.*, **19**, 22–68.

European Parliament (2011) The use of gender in insurance pricing, Directorate General for Internal Policies, European Parliament, Brussels. (Available from `http://www.europarl.europa.eu/RegData/etudes/note/join/2011/453214/IPOL-FEMM_NT(2011)453214_EN.pdf`.)

European Union Council (2004) Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. European Union Council, Brussels. (Available from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:373:0037:0043:en:PDF.)

Fair, W. (1979) Hearings before the Subcommittee on Consumer Affairs of the Committee on Banking, Housing and Urban Affairs, United States Senate, 96th Congress, 1st Session on S. 15, June 4 and 5, 1979. US Government Printing Office, Washington DC.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. (2017) Predictably unequal?: The effects of machine learning on credit markets. (Available from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3072038.)

Greenacre, M. (2016) *Correspondence Analysis in Practice*, 3rd edn. Barcelona: Chapman and Hall–CRC.

Hand, D. J. (1998) Consumer credit and statistics. In *Statistics in Finance* (eds D. J. Hand and S. D. Jacka). London: Arnold.

Hand, D. J. (2009) Modern statistics: the myth and the magic. *J. R. Statist. Soc.* A, **172**, 287–306.

Hand, D. J. (2012a) Credit scoring, insurance and discrimination. In *Statistics, Science, and Public Policy XVI: Risks, Rights, and Regulations* (ed. A. M. Herzberg), pp. 85–90.

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

Hand, D. J. and Henley, W. E. (1997) Statistical classification methods in consumer credit scoring: a review. *J. R. Statist. Soc.* A, **160**, 523–541.

Hanley, J. A. and McNeil, B. J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Johnson, R. W. (2004) Legal, social and economic issues in implementing scoring in the United States. In *Readings in Credit Scoring* (eds L. C. Thomas, D. B. Edelman and J. N. Cook). Oxford: Oxford University Press.

Jung, K. M. and Thomas, L. C. (2008) A note on coarse classifying in acceptance scorecards. *J. Oper. Res. Soc.*, **59**, 714–718.

Kohavi, R. (1995) A study of cross-validation and boot-strap for accuracy estimation and model selection. In *Proc. 14th Int. Jt Conf. Artificial Intelligence*, vol. 2, pp. 1137–1143. San Francisco: Morgan Kaufmann.

Korb, K. and Nicholson, A. E. (2010) *Bayesian Artificial Intelligence*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Lin, S.-M., Ansell, J. and Andreeva, G. (2012) Predicting default of a small business using different definitions of financial distress. *J. Oper. Res. Soc.*, **63**, 539–548.

McDonald, S. (2015) Indirect gender discrimination and the 'Test-Achats Ruling': an examination of the UK motor insurance market. *Royal Economic Society Conf., Manchester*. (Available from https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=RES2015&paper_id=791.)

Oxera (2010) The use of Gender in insurance pricing: analysing the impact of a potential ban on the use of gender as a rating factor. *Research Paper 24*. ABI. (Available from http://www.oxera.com/Oxera/media/Oxera/The-use-of-gender-in-insurance-pricing.pdf?ext=.pdf.)

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann.

Phelps, E. S. (1972) The statistical theory of racism and sexism. *Am. Econ. Rev.*, **62**, 659–661.

R Development Core Team (2010) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Royal Statistical Society (2017) Written evidence submitted by The Royal Statistical Society (ALG0071) to 'Algorithms in decision-making' inquiry. Royal Statistical Society, London. (Available from http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/science-and-technology-committee/algorithms-in-decisionmaking/written/69144.html.)

Schreiner, M. (2002) Scoring: the next breakthrough in micro-credit. Microfinance Risk Management, Centre for Social Development, Washington University, St Louis.

Science and Technology Committee (2017) Algorithms in decision-making inquiry. Science and Technology Committee, London. (Available from https://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/inquiry9/.)

Scutari, M. (2010) Learning Bayesian networks with the bnlearn R package. *J. Statist. Softwr.*, **35**, 1–22.

Siddiqi, N. (2006) *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Hoboken: Wiley.

Sourial, N., Wolfson, C., Zhu, B., Quail, J., Fletcher, J., Karunananthan, S., Bandeen-Roche, K., Béland, F. and Bergman, H. (2010) Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *J. Clin. Epidem.*, **63**, 638–646.

Stiglitz, J. E. (1993) Approaches to the economics of discrimination. *Am. Econ. Rev.*, **62**, 287–295.

Taylor, W. (2011) Proving racial discrimination and monitoring fair lending compliance: the missing data problem in nonmortgage credit. *Rev. Bankng Finan. Law*, **31**, 199–264.

Thomas, L. C., Edelman, D. B. and Crook, J. N. (2017) *Credit Scoring and Its Applications*, 2nd edn. Philadelphia: Society for Industrial and Applied Mathematics Publishing.

Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006) The max-min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.*, **65**, 31–78.

US Government (2015) Home Mortgage Disclosure (Regulation C), 66128 Federal Register/Vol. 80, No. 208/Wednesday, October 28, 2015 / Rules and Regulations. US Government, Washington DC. (Available from `https://www.gpo.gov/fdsys/pkg/FR-2015-10-28/pdf/2015-26607.pdf`.)